# Exam PA April 2026 Project Statement

## General Information for Candidates

This examination has 10 tasks numbered 1 through 10 with a total of 70 points. The points for each task are indicated at the beginning of the task, and the points for subtasks are shown with each subtask.

Each task pertains to the business problem described below. We recommend that you read the business problem and data dictionary to learn additional context about each task. Additional information on the business problem may be included in specific tasks—where additional information is provided, including variations in the target variable, it applies only to that task and not to other tasks. You may use Excel for calculation for any of the tasks, but all answers must be submitted in the Word document. *If you upload the Excel document, it will not be looked at by the graders.* Neither R nor RStudio are available.

The responses to each specific subtask should be written after the subtask and the answer label, which is typically ANSWER, in this Word document. Each subtask will be graded individually, so be sure any work that addresses a given subtask is done in the space provided for that subtask. Some subtasks have multiple labels for answers where multiple items are asked for—each answer label should have an answer after it.

Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. The answer should be confined to the question as set. No response to any task needs to be written as a formal report. Unless a subtask specifies otherwise, the audience for the responses is the examination grading team and technical language can be used.

Prior to uploading your Word file, it should be saved and renamed with your five-digit candidate number in the file name. If any part of your exam was answered in French, also include “French” in the file name. Please keep the exam date as part of the file name.

## Business Problem

*You work for an actuarial consulting firm that applies predictive modeling techniques to business problems in Los Angeles County, U.S. Los Angeles County has a densely populated inner city with a low-income population and less populous suburbs, which generally have higher-income populations.*

*You are currently working on projects to bring products and services to people in Los Angeles County who do not own cars. To support your work, you are using a dataset with census, survey, and geospatial data for each census tract in Los Angeles County.*[1]

***We recommend that you review the data dictionary to see additional information about each variable.***

[1] County of Los Angeles Open Data

## Data Dictionary

| Variable | Data Type / Range / Example | Description |
|---|---|---|
| tract | Character<br>Example: 0637113231 | Unique identifier for census tract. Each tract is a subdivision of Los Angeles County. |
| no_vehicle | Numeric<br>Range: 0 - 1,268 | Number of households surveyed which have no vehicle. |
| no_vehicle_universe | Numeric<br>Range: 29 - 5,202 | Total number of households surveyed. |
| no_vehicle_pct | Numeric<br>Range: 0 - 82.1 | Percentage of households with no vehicle: 100 * no_vehicle / no_vehicle_universe |
| sup_dist | Character<br>Examples: District 1, District 2 | Supervisorial District. Each district has a representative on the Los Angeles County Board of Supervisors |
| csa | Character<br>Examples: Los Angeles - Hollywood, City of Inglewood | Combined Statistical Area (CSA). A CSA is a population center and adjacent communities. |
| spa | Character<br>Examples: SPA 2 - San Fernando, SPA 6 - South | Service Planning Area. SPAs are used in providing health services to meet the needs of each area. |
| Shape__Area | Numeric<br>Range: 48,365 - 16,086,908,851 | Land area of the census tract in square feet. |
| Shape__Length | Numeric<br>Range: 2,815 - 915,243 | Perimeter of the census tract in feet. |
| percent_hs | Numeric<br>Range: 30.1 - 100 | Percentage of people in the tract who have graduated high school. |
| percent_bachelors | Numeric<br>Range: 1.3 - 91.7 | Percentage of people in the tract who have a bachelor's degree. |
| disabled_pct | Numeric<br>Range: 1.1 - 56 | Percentage of people in the tract who are disabled. |
| income_percap | Numeric<br>Range: 9,284 - 214,838 | Income per capita for the tract. |
| med_hh_income | Numeric<br>Range: 7,000 - 250,001 | Median Household Income in the census tract. Values greater than $250,000 are entered as $250,001 |

| | | |
|---|---|---|
| ami_category | Character<br>Examples: Low Income, Moderate Income | Median Household Income category. |

## Task 1 (*7 points*)

Your actuarial consulting firm has been approached by an economic development organization that is looking to invest in various parts of Los Angeles. They are considering options that enhance the walkability of local communities and are looking for recommendations on potential areas in which to invest.

**(a)** (*1 point*) List four common characteristics of a predictive modeling problem.

*Candidate performance was mixed on this task. To earn full-credit, candidates needed to provide 4 distinct valid elements from the source materials. Paraphrased responses (KPIs, Quality data) were also accepted as alternative answers if candidate provided supporting comments. Most candidates addressed the need for data to do predictive modeling but failed to recall other key elements from the learning modules; partial points were given for those answers. No points were awarded to candidates who listed components of GLM or Tree models, since the task is asking for common characters of predictive modeling, not of specific types of models.*

**ANSWER:**

*The candidate must list four of the following elements:*

- You can clearly identify and define a business issue that needs to be addressed.
- You can address the issue with a few well-defined questions.
- You have lots of good and useful data that can be used to answer these questions.
- You are reasonably certain the predictions will drive actions or increase understanding.
- You are reasonably certain it is better than any existing approach.
- You can continue to monitor and update the models as new data comes in.

---

Your analyst has developed the following table in support of the client request:

| District | Count of Census Tracts | No Vehicle Count | No Vehicle Percentage | Disabled Count | Disabled Percentage |
|---|---|---|---|---|---|
| District 1 | 497 | 68,534 | 11.1% | 214,398 | 11.2% |
| District 2 | 506 | 71,263 | 10.8% | 223,645 | 11.4% |
| District 3 | 557 | 63,142 | 8.3% | 211,960 | 10.5% |
| District 4 | 487 | 44,878 | 6.8% | 210,634 | 10.4% |
| District 5 | 448 | 39,990 | 6.1% | 203,690 | 11.0% |

**(b)** (*2 points*) Identify the area of predictive analytics (descriptive, predictive, or prescriptive analytics) that the table represents and explain whether it addresses the client's needs. Do not directly discuss the data presented in the table.

*Candidate performance was mixed on this task. To earn full-credit, candidates needed to identify that the table represents descriptive analytics while the client's request is prescriptive, and provide valid justifications to evaluate whether this table meets the client's needs. Most candidates received points on correctly identifying descriptive analytics but failed to connect to the business problem.*

**ANSWER:**

The table represents descriptive analytics – It gives insight into what happened in the past. The client's request is a prescriptive request – what if we build in this area? The table does not meet the client's needs because it merely shows what has happened and provides no information on where they should focus on development.

---

The client has also provided you with survey information from residents of various neighborhoods being considered for potential development opportunities. Included in this survey information are additional variables capturing job codes, home values, marital status, and religious status.

They have recommended that you build models that include all this additional information.

**(c)** (*2 points*) Critique your client's recommendation of including all the additional variables in your model.

*Candidate performance was mixed on this task. To earn full credit, candidates needed to consider the bias/variance tradeoff in their critique and address the impacts on predictive models. Partial credit was awarded to candidates who critiqued those added variables from the perspective of personally identifiable information, data sensitivity, or the ethics of usage. Candidates are reminded that "critique" in the SOA Verb List requires discussion of both strengths and weaknesses to receive full credit. While the model solution below discusses overfitting and high variance as a weakness, an accepted alternative answer was the granularity mismatch of the proposed variables.*

**ANSWER:**

The additional variables can help identify other factors that impact vehicle status and better inform development opportunities. However, the additional variables may cause overfitting when used together with variables that are already in the dataset, leading to high variance in the model.

---

The survey includes a comment box where residents can enter suggestions for future projects.

**(d)** (*2 points*) Identify what kind of data these comments represent and state the pros and cons of including this in your analysis.

*Candidates performed very well on this task. To earn full credit, candidates needed to describe the data as unstructured and have valid pros and cons.*

**ANSWER:**

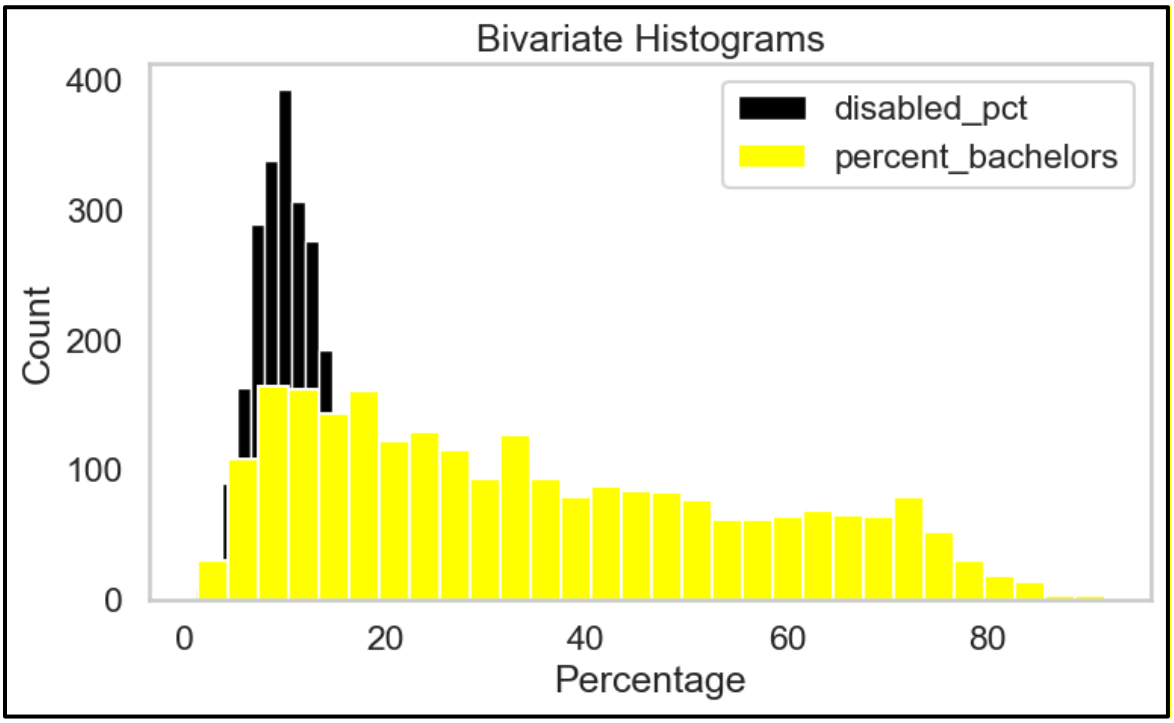
The data in the comment box is unstructured data. Suggestions from residents may contain useful data for our analysis that cannot be naturally fit into tabular form, and we could extract useful information

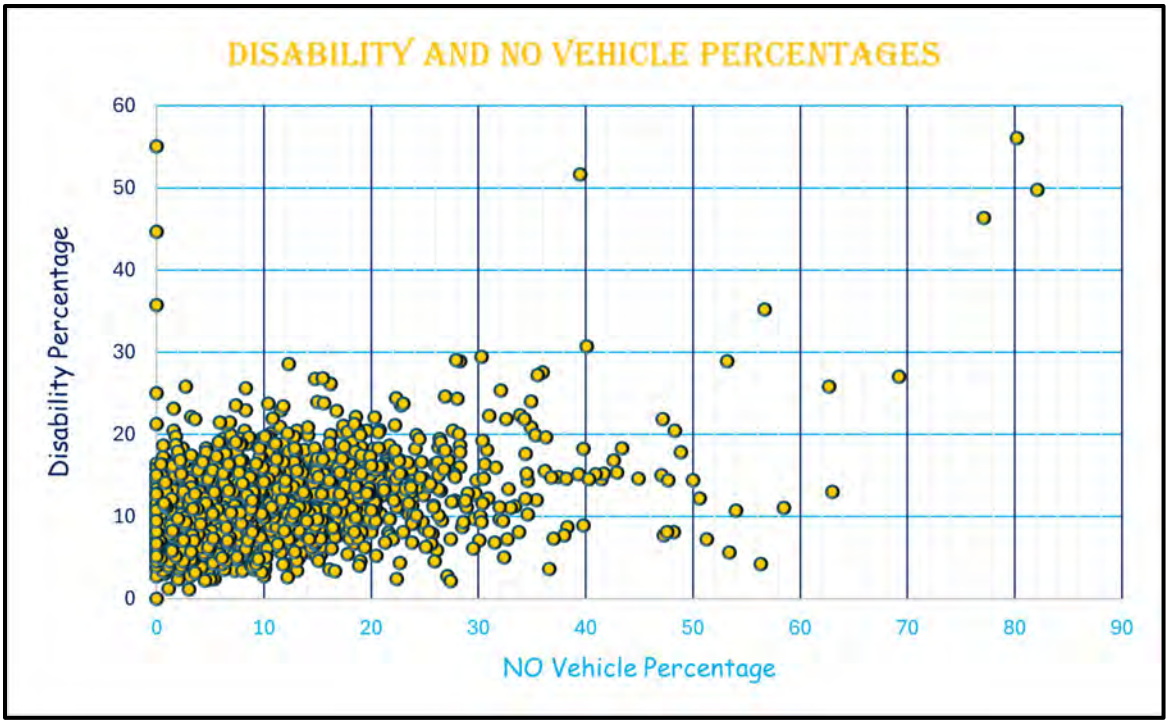
and generate features from this data. However, the feature generation process will require a decision of which and how many features are appropriate for building a parsimonious model.

---

### Task 2 (*10 points*)

You have asked your assistant to develop bivariate or multi-variate visualizations connecting the rate of vehicle ownership with either education level or disability. They have provided the following graphs as part of this assignment:



**(a)** (*3 points*) Critique each graph with regard to data visualization design guidelines, identifying at least two good and at least two bad elements in each graph.

*Candidates performed poorly on this task. Many candidates analyzed the content of the graphs instead of critiquing the design choices. Candidates who critiqued the graph design frequently only identified the bad elements.*

**ANSWER:**

| | Good Elements | Bad Elements |
|---|---|---|
| Graph 1 | Labeled Axes<br>Contrasting Colors<br>Inclusion of a Legend | Different Widths to the Histograms<br>Overlapping representation |
| Graph 2 | Labeled Axes<br>Minor Gridlines to help with value | Contrasting Fonts<br>Mixed Colors |

---

**(b)** (*2 points*) Critique the two graphs with regard to the analysis you requested.

*Candidates generally struggled with this task because they often evaluated the graphs in isolation rather than assessing how well each visualization addressed the original request. Common errors included failing to recognize the univariate nature of Graph 1, missing the inverse relationship shown in Graph 2, or limiting their critique to only one graph. Full credit was earned by candidates who clearly explained why Graph 1 was not appropriate for the requested analysis, distinguished between the univariate and bivariate visualizations, and provided specific critiques of both graphs in the context of the question.*

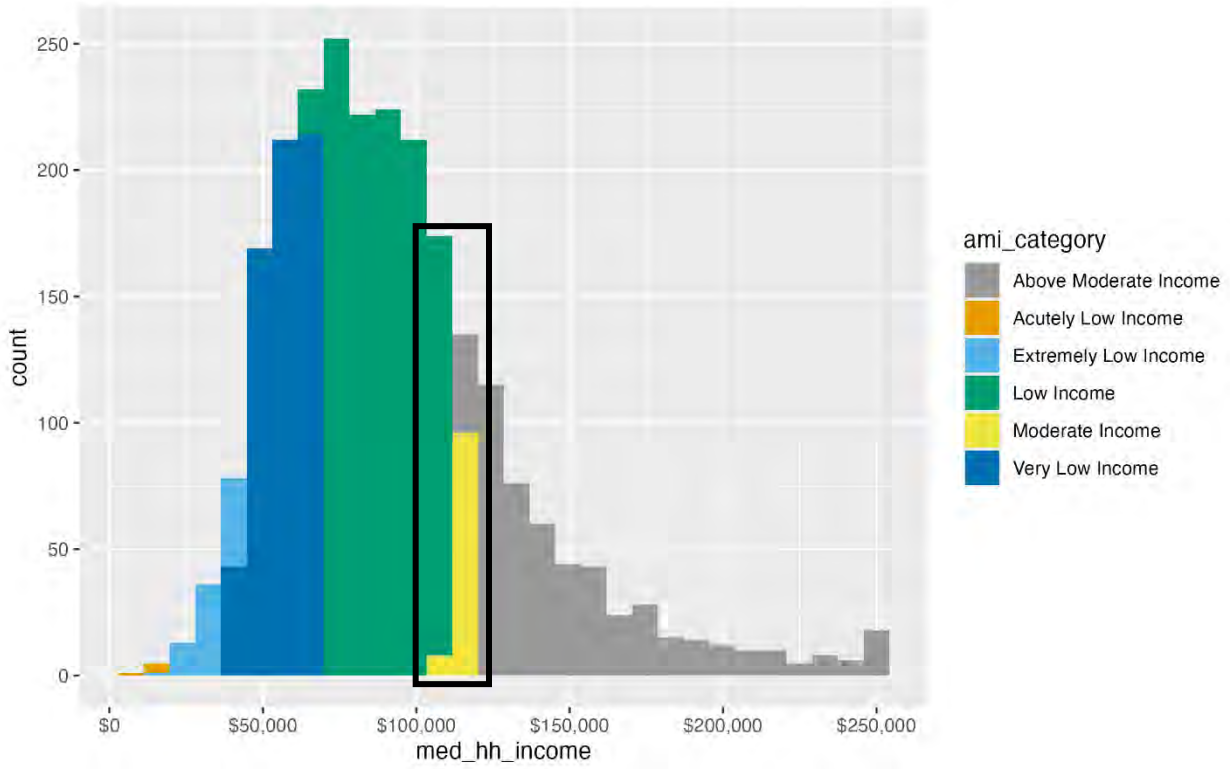
**ANSWER:**

The first graph, while labeled bivariate, is in fact just univariate analysis since it only shows the distribution of each variable. A bivariate analysis shows the relationship between two variables.

The second graph does show a bivariate analysis connecting vehicle percentage with disability percentage. The scatterplot is helpful for understanding the relationship between these two variables.

---

You want to understand the income-related variables in the dataset better and ask your assistant to investigate the relationship between **med_hh_income** and **ami_category**. They create the histogram below:




**(c)** (*3 points*) Interpret the following aspects of the histogram:

i. The overall relationship between these two variables

ii. The section of the histogram boxed in black, where there are bars for low, moderate, and above moderate income

iii. The relatively high density around $250,000

*Candidates generally performed better at describing what they observed in the visualizations than interpreting the underlying data characteristics and relationships. Common challenges included failing to move beyond basic descriptions of distribution shapes or categories, misinterpreting bin ranges and category labels, and overlooking more nuanced concepts such as positive correlation, data truncation, and the implications for frequency counts. Full credit was earned by candidates who provided detailed and accurate interpretations across all parts of the subtasks, clearly explaining the key findings and their significance.*

**ANSWER:**

i. The chart shows a strong relationship between **med_hh_income** and **ami_category**. There is clear stratification with little overlap by category.

ii. The section in the box shows that the "Moderate Income" **ami_category** corresponds to a very narrow range of values of **med_hh_income**, and there are relatively few observations for this category compared to "Low Income" and "Above Moderate Income."

iii. The higher density around $250,000 is a result of the data being censored at $250,000 as indicated in the data dictionary.

---

**(d)** (*2 points*) Explain the problems that may arise from using both variables in a GLM, including considerations of model interpretability and performance.

*Candidates generally performed well on this subtask, with most identifying collinearity as the issue. While collinearity by definition implies two numeric variables, the principle is the same in that these two variables are essentially measuring the same thing. This is a fine distinction. and the use of the term collinearity was accepted. However, many responses lacked depth, often restating the definition of collinearity without adequately explaining its mechanism or its impact on both predictive performance and interpretability in a GLM. Full credit was earned by candidates who connected the statistical concept to its practical consequences and clearly addressed all aspects of the question.*

**ANSWER:**

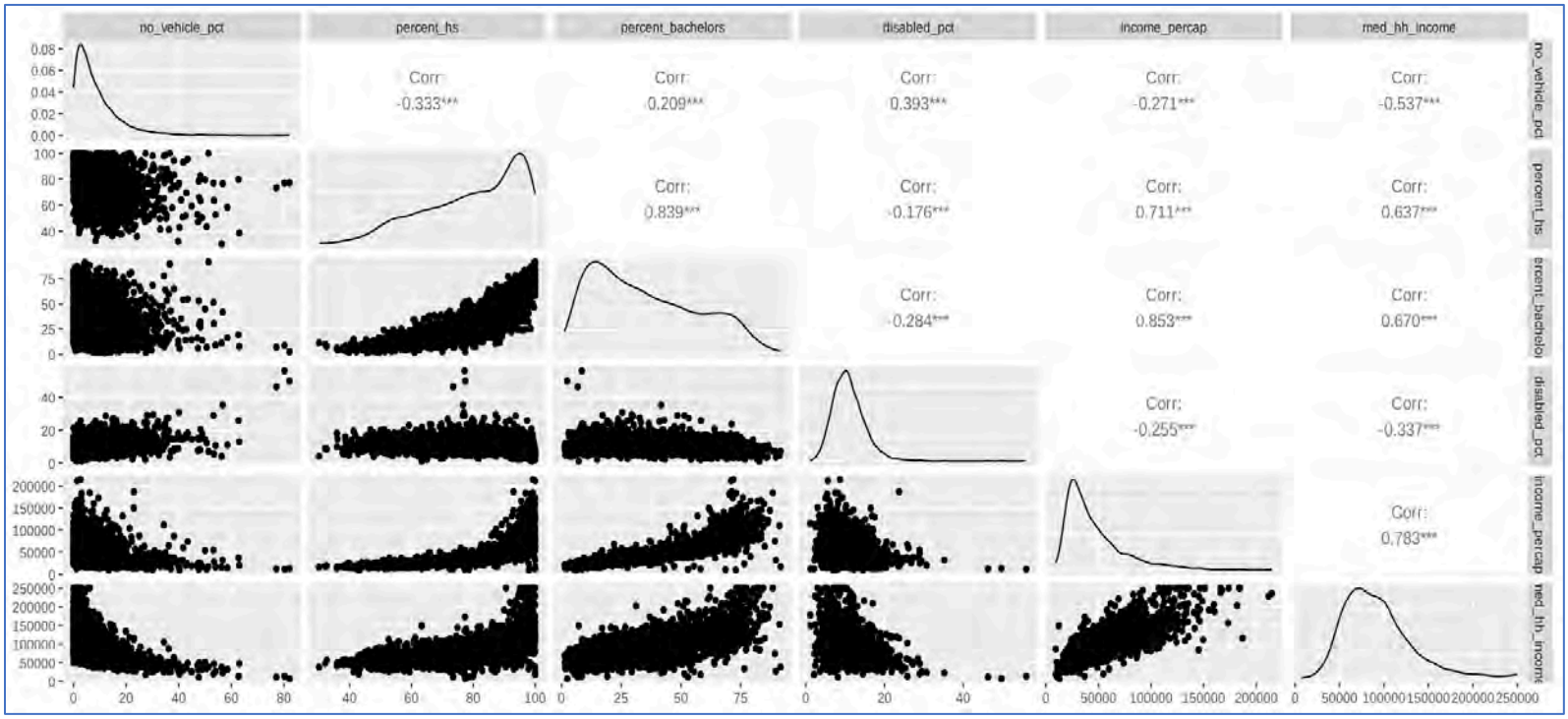
Both variables should not be included in a GLM. Their strong relationship means that the model would likely suffer from incorporating two variables that measure the same thing (essentially collinearity). This inflates standard errors, making coefficients unstable. This by itself lowers interpretability; in addition, the similar-sounding variables may be challenging for stakeholders.

---

## Task 3 (*6 points*)

The Los Angeles County Metropolitan Transportation Authority has asked your firm to cluster census tracts into a small number of neighborhood types to identify transit-vulnerable area clusters. *Transit-vulnerable* is defined as having low socioeconomic status, having poor access to private vehicles, and being more reliant on public transit.

You are provided with a pairs plot along with correlations of the variables selected for this study.




**(a)** (*2 points*) Explain the importance of standardizing these six variables before fitting a K-means model.

*In general, candidates did well in this subtask. The most common issue was that candidates provided overly generic responses that were not clearly tied to the specifics of the problem. In particular, many failed to explicitly mention Euclidean distance or make reference to specific variables in the question.*

*Full-credit responses required candidates to clearly articulate that K-means clustering relies on Euclidean distance and to explain that, without standardization, variables with larger magnitudes will dominate the distance calculation. In addition, candidates needed to explicitly reference specific variables from the problem. Partial credit responses generally reflected incomplete elements for full credits.*

**ANSWER:**

*K*-means uses Euclidean distances between observations for clustering. The percentage variables and income variables are on very different scales. Without standardization, income variables would

dominate distances. Standardizing makes variables comparable and prevents one variable from dominating.

---

You fit a 3-cluster *K*-means model on the standardized variables. The table below shows cluster means on the original scale.

| cluster_k3 | n_tracts | no_vehicle_pct | percent_hs | percent_bachelors | disabled_pct | income_percap | med_hh_income |
|---|---|---|---|---|---|---|---|
| <fct> | <int> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 890 | 5.461124 | 94.71326 | 59.23528 | 9.140674 | 73369.86 | 127711.58 |
| 2 | 367 | 23.487738 | 68.03515 | 22.07466 | 15.867847 | 27583.96 | 54151.57 |
| 3 | 1196 | 6.646488 | 72.57333 | 21.04933 | 10.791388 | 30689.10 | 80167.38 |

**(b)** (*2 points*) Identify which cluster is most transit-vulnerable. Describe the characteristics of the chosen cluster.

*Overall, candidates performed strongly on this question, with most correctly identifying the appropriate cluster (Cluster 2). However, the primary gap was in the depth of justification. Some candidates failed to reference specific variables altogether, while others cited only one or two variables instead of providing a more comprehensive explanation covering more key variables aligned with the definition of transit vulnerability.*

*Full credit responses required candidates to correctly identify Cluster 2 and provide a well supported justification using multiple relevant variables. In particular, candidates needed to clearly explain how these variables compare to other clusters and support the classification. Partial credit was given to responses that identified Cluster 2 but with incomplete justification.*

**ANSWER:**

Cluster 2 is the most transit-vulnerable.

- Highest no_vehicle_pct: most households without cars across the 3 clusters
- Lowest incomes: least ability to purchase vehicles or other transport
- Lowest education and highest disabled_pct: more likely to rely on public transit

---

Your assistant proposes adding **spa**, a categorical variable with 8 levels, to the clustering analysis using the following approach:

- Create seven one-hot variables to replace **spa**: **spa_1**, **spa_2**, …, **spa_7**.
- Standardize all variables.
- Fit *K*-means using the original standardized variables plus the seven standardized variables **spa_1**, **spa_2**, … **spa_7**.

Assume K-means uses Euclidean distance on the standardized predictors.

**(c)** (*2 points*) Critique your assistant's approach of adding the categorical variable, including considerations of how the Euclidean distance calculation will be impacted.

*Candidates generally struggled with this question, often providing vague or generic responses that demonstrated an incomplete understanding of one-hot encoding and its impact on clustering. Common issues included failing to recognize that high-cardinality categorical variables increase dimensionality, overlooking how one-hot encoded features can dominate Euclidean distance calculations, and not explaining that clustering may become driven primarily by the new variable rather than other important features. Full-credit responses clearly described the one-hot encoding process, its effect on distance calculations and dimensionality, and the potential for distorted clustering outcomes due to the disproportionate influence of the added variable.*

**ANSWER:**

One-hot encoding avoids arbitrary ordering, but 7 more standardized spa variables are created in addition to the 6 original variables. They will all contribute to the distance calculation. However, with 7 standardized spa variables added, they can dominate the distance calculation, causing clusters to align mostly with region rather than with original socio-economic variables

---

## Task 4 (*7 points*)

You would like to understand the age distribution of residents across different census tracts. You begin by applying principal components analysis (PCA) to population age data to get more informative variables.

**(a)** (*2 points*) Compare and contrast *K*-Means Clustering and PCA as unsupervised learning approaches.

*Candidate performance on this question was mixed, as many responses failed to fully address both the comparison and contrast requirements, often emphasizing one while neglecting the other. Common responses included providing only superficial similarities, focusing disproportionately on differences, or describing PCA and K-means independently without explicitly relating them. Full-credit responses clearly articulated both a meaningful similarity and a meaningful difference, including how the methods operate and the distinct outputs they produce.*

**ANSWER:**

Compare: *K*-means Clustering and PCA are both unsupervised learning methods that seek to gain useful information about data without having a specific target value to predict.

Contrast: The two methods have different outputs. While PCA results in dimension reduction, K-means clustering seeks to group similar data points together to reduce the number of data points.

PCA seeks to embed the entries in a dataset in a lower-dimensional space to reduce the complexity of the data and provide more directly informative features for use in downstream supervised learning, visualization, or imputation. In PCA, each data point is represented by orthogonal transformations of the original variables.

---

The data relating to age is split into 14 age buckets, each with a count of the number of individuals. **POP22_AGE_0_4** is the count of people between age 0 and 4, **POP22_AGE_5_9** is the count of people between age 5 and 9, etc. A sample from the data is shown here for reference where each row represents one census tract:

| POP22_ AGE_0_4 | POP22_ AGE_5_9 | POP22_ AGE_10_14 | POP22_ AGE_15_17 | POP22_ AGE_18_19 | Table continues… |
|---|---|---|---|---|---|
| 209 | 251 | 219 | 130 | 90 | |
| 206 | 143 | 174 | 97 | 88 | |
| 131 | 148 | 130 | 106 | 65 | |
| 158 | 199 | 205 | 102 | 68 | |

You perform PCA on the standardized data and receive the following coefficients for the first three principal components:

| Column Name | PC1 | PC2 | PC3 |
|---|---|---|---|
| POP22_AGE_0_4 | 0.31 | -0.17 | -0.11 |
| POP22_AGE_5_9 | 0.30 | -0.16 | -0.01 |
| POP22_AGE_10_14 | 0.30 | -0.16 | 0.04 |
| POP22_AGE_15_17 | 0.29 | -0.14 | 0.08 |
| POP22_AGE_18_19 | 0.10 | -0.27 | 0.65 |
| POP22_AGE_20_24 | 0.20 | -0.33 | 0.44 |
| POP22_AGE_25_29 | 0.27 | -0.21 | -0.26 |
| POP22_AGE_30_34 | 0.25 | -0.13 | -0.38 |
| POP22_AGE_35_44 | 0.30 | -0.05 | -0.27 |
| POP22_AGE_45_54 | 0.32 | 0.06 | -0.05 |
| POP22_AGE_55_64 | 0.31 | 0.19 | 0.04 |
| POP22_AGE_65_74 | 0.28 | 0.35 | 0.10 |
| POP22_AGE_75_84 | 0.24 | 0.48 | 0.16 |
| POP22_AGE_85_100 | 0.19 | 0.51 | 0.17 |

Your assistant suggests that since all the coefficients in PC1 are positive, we can remove PC1 from our analysis and proceed with PC2, PC3.

**(b)** (*2 points*) Critique your assistant's statement.

*Candidates performed well on this question, with most correctly recognizing that PC1 explains the greatest proportion of variance and therefore should not be removed. The most common weakness was failing to explain the meaning of all-positive coefficients in PC1, while some candidates lost credit by focusing on irrelevant concepts such as sign flipping or rerunning PCA rather than addressing the interpretation of the component and its variance contribution.*

**ANSWER:**

This component appears to be a general scale vector that represents the overall size of a given census tract. All coefficients in the vector increase with increases to any age bucket. It is important to include it because as the first principal component, it explains the greatest amount of variance within the dataset.

---

You are provided with the following information about the **POP22_AGE_0_4** variable:

| | |
|---|---|
| Mean | 158.835 |
| Variance | 9518.094 |

**(c)** (*3 points*) Using the PC coefficients above, estimate **POP22_AGE_0_4** for a point with the following weights in the original, unstandardized units:

| | |
|---|---|
| PC1 Weight | 2.194 |
| PC2 Weight | 1.234 |

| | |
|---|---|
| PC3 Weight | 0.099 |

Show your work.

*Overall, candidates performed poorly on this question, with many struggling to correctly apply PCA coefficients and weights and to reverse the standardization process. While some candidates successfully completed the initial weighted calculation, they often failed to properly unstandardize the result by reversing the scaling and centering, leading to incomplete or incorrect solutions. Full-credit responses correctly performed both calculation steps and arrived at the final estimate, whereas partial-credit responses reflected only a partial understanding of the required process.*

**ANSWER:**

Begin by multiplying the coefficient for each of the PCs with its respective weight and adding the results:

Scaled and Centered Estimate = (PC1[POP22_AGE_0_4] * PC1 Weight) + …

= (2.194 * 0.31) + (1.234 * -0.17) + (0.099 * -0.11)

= 0.45947

Then reverse the scaling and centering performed on the original dataset:

0.445 * (9518.094) ^ 0.5 + 158.835

= 203.66

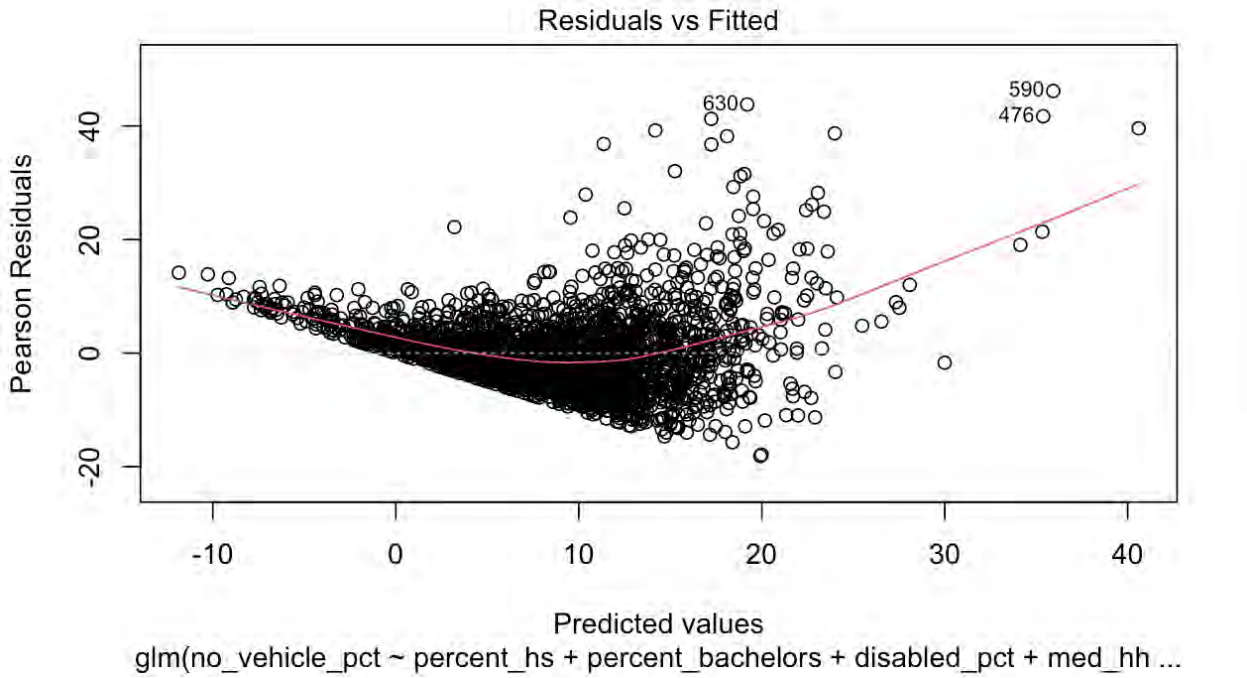
---

## Task 5 (*5 points*)

Your manager asks you to build a linear model to predict car ownership using demographic and economic variables. You fit the following model:

```
m0 <- glm(
  no_vehicle_pct ~ percent_hs + percent_bachelors + disabled_pct
                 + med_hh_income,
  family = gaussian(link = 'identity'),
  data = mdat,
  na.action = na.omit
)
```

A residual plot for this model is shown below:




**(a)** (*2 points*) Critique the choice of the Gaussian family with identity link for modeling **no_vehicle_pct**.

*Candidates generally performed well on this question, with most correctly identifying heteroscedasticity and recognizing that a Gaussian model was inappropriate due to its unbounded nature. A small number of candidates incorrectly concluded that the Gaussian distribution was appropriate, and while most recognized the residuals appeared approximately normal, fewer explicitly discussed the presence of negative values in the residual plot and their implications.*

**ANSWER:**

The residual plot shows that two assumptions of the Gaussian family are violated:

1) The response is unbounded and approximately normal. We know that the response variable is bounded by 0 and 100. Although in certain situations a Gaussian family may make reasonable predictions of a bounded variable, we can see that many predicted values are between 0, with some near -10.

2) Variance is constant across observations. The residual plot shows heteroscedasticity, with variance highest for predicted values around 20 and lower variance elsewhere.

These violations suggest the Gaussian distribution with identity link is inappropriate.

---

**(b)** (*2 points*) Recommend a more appropriate GLM family and link function. Justify your recommendation based on the plot and characteristics of the response variable.

*Candidates generally recognized that a model with a lower-bounded distribution was needed, with many recommending a gamma distribution because it can accommodate positive, right-skewed outcomes. However, a few candidates recommended a binomial family or recognized the need to transform the target by dividing by 100. Many provided incomplete reasoning, often citing model benefits such as handling skewness, ensuring positive predictions and ease of interpretation with a log link.*

**ANSWER:**

I recommend a binomial family with logit link (predicting **no_vehicle_pct** / 100, a version of the variable bounded by 0 and 1). The binomial distribution is appropriate for response variables that represent a proportion or probability. This distribution ensures that the predictions will be in the valid range. Having a more appropriate range and distribution will also likely result in a more reasonable distribution of residuals.

---

**(c)** (*1 point*) Explain briefly how interpretation of coefficients changes under your recommended GLM family and link function compared to the GLM with Gaussian family and identity link.

*Candidates generally performed well on this question, with most earning full credit by providing a recommendation and correctly explaining how the choice of family or link function would affect coefficient interpretation. Weaker responses focused on the ease of interpretation rather than the impact on the coefficients themselves, and some candidates received partial credit for discussing only their recommended family without comparing it to the Gaussian model.*

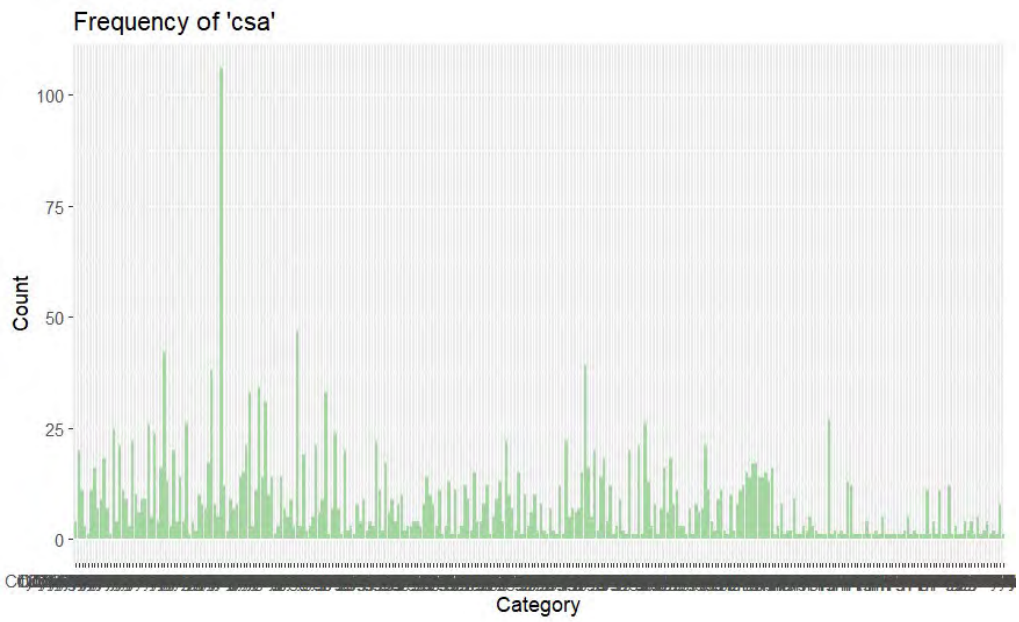
**ANSWER:**

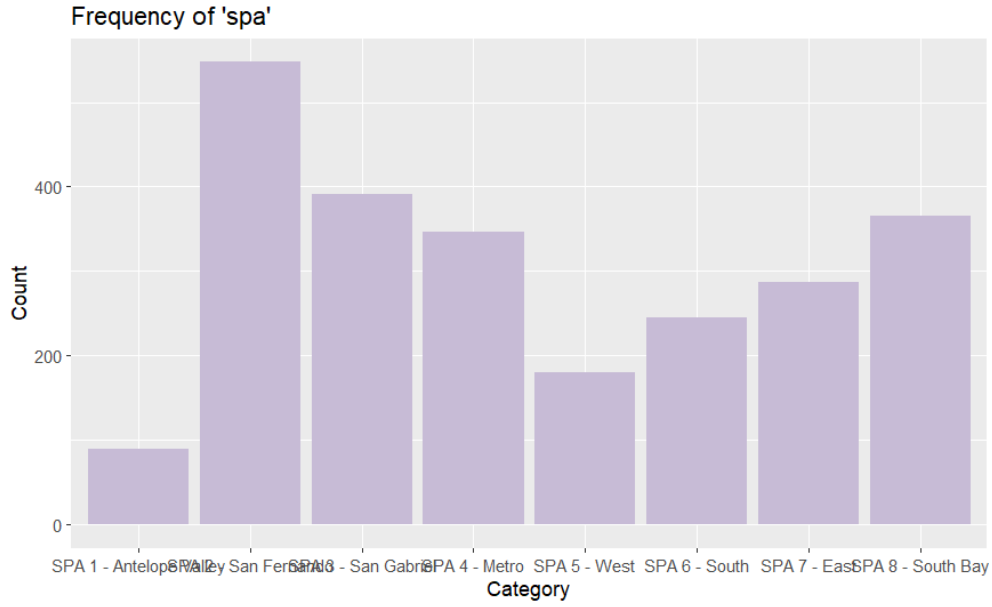
Coefficients for a binomial/logit GLM represent multiplicative change in log-odds of **no_vehicle_pct** for a one-unit increase in the predictor, holding all other variables constant. Coefficients for the Gaussian/identity GLM directly represent the change in percentage points for a one-unit increase in the predictor, holding other variables constant.

---

## Task 6 (*7 points*)

Your assistant is investigating geographic data to understand how it could be used in models to predict **no_vehicle_pct**. They produce the following bar charts for the geographic factor variables **csa** and **spa**:



**(a)** (*2 points*) Describe one advantage and one disadvantage of using each variable as a predictor in a model.

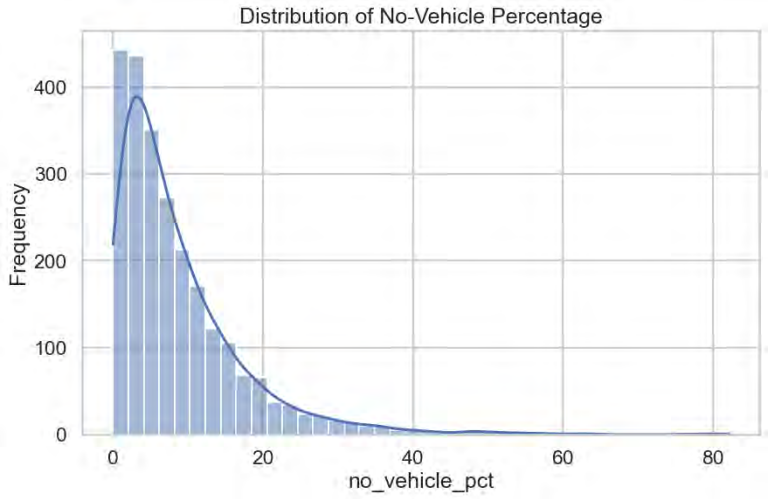
*Candidates frequently struggled with this question by focusing on the visualizations themselves rather than discussing how the variables affected model performance or interpretability. Many responses identified only disadvantages rather than advantages. Other candidates failed to address challenges associated with high-cardinality categorical variables (such as sparse categories, binarization, or data imbalance) or did not explicitly connect their points to the model. Full-credit responses provided both an advantage and a disadvantage for each variable and clearly explained the implications for the model, whereas partial-credit responses were incomplete, vague, or lacked a direct link to model outcomes.*

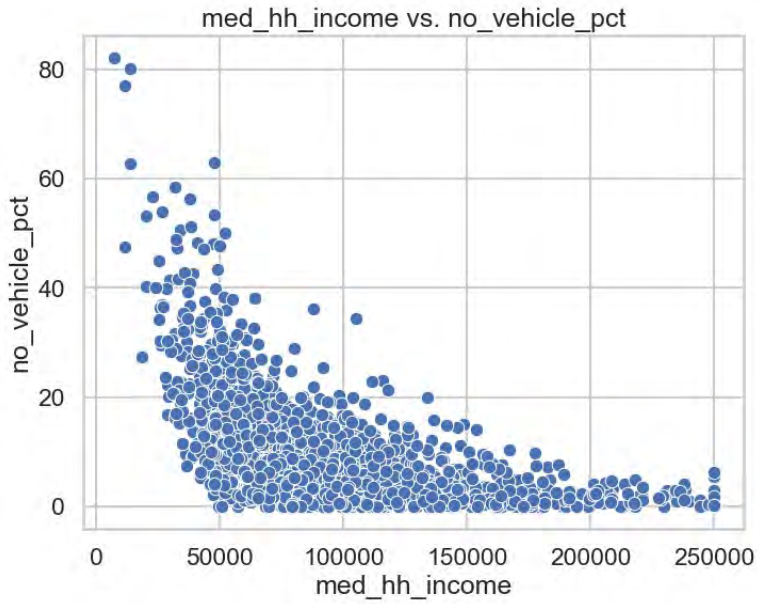
**ANSWER:**

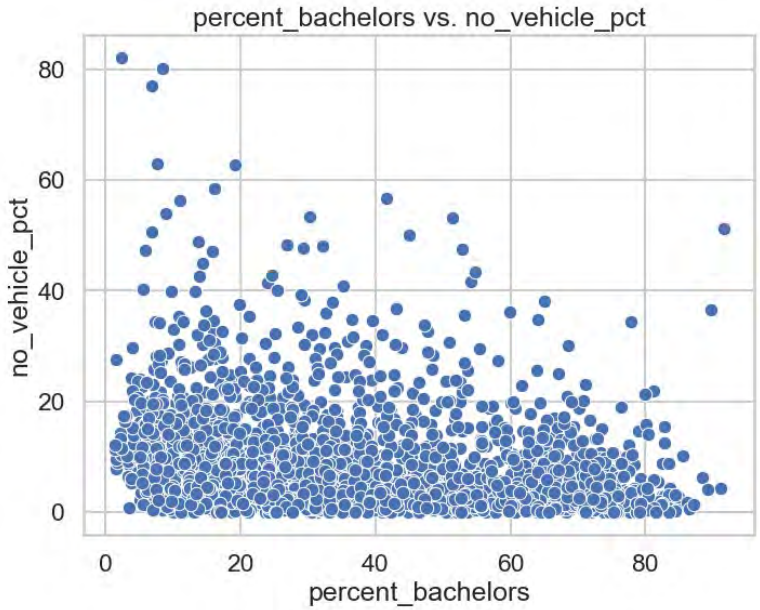
| | Advantage | Disadvantage |
|---|---|---|
| **csa** | This variable has many levels, which can be very predictive in certain model types, including random forest and boosted trees | The high dimensionality can cause overfitting and interpretability problems in other model types, like GLM |
| **spa** | Each level of the variable is interpretable (e.g. the name is descriptive) and has enough observations to be useful in most models | There are relatively few observations in the SPA 1 category. Model stability may improve by combining this with another category |

---

After reviewing the geographic variables, you ask your assistant to review the distribution of the response variable, **no_vehicle_pct,** and perform bivariate analysis on the remaining numeric variables. Your assistant has produced additional statistical summaries and charts to support the analysis:






**(b)** (*3 points*) Describe the distribution of **no_vehicle_pct** based on the graphs above and interpret the relationship between this target variable, **med_hh_income**, and **percent_bachelors**.

*Most candidates correctly identified the skewness of the distribution. However, many missed important features such as the thin tail, the specific drop-off point, outliers, and the non-linear nature of the relationship, often defaulting to a linear interpretation. Full-credit responses went beyond identifying basic patterns by discussing the non-linear trend, outliers, and business implications of the observed relationships, while partial-credit responses omitted one or more of these key elements.*

**ANSWER:**

**no_vehicle_pct** – The distribution is right-skewed with a thin tail. Most of tracts have low vehicle non-ownership and some smaller subset of tracts have very high more than 20% non-ownership percentage.

**med_hh_income** – The income graph shows a strong downward trend with vehicle ownership. As the income rises, the percentage of households without vehicles drops. The relationship is not linear and there is a significant drop in households without vehicles when income is below around $40,000.

**percent_bachelors** – This graph shows a downward trend, but the relationship has more noise than **med_hh_income**. There also appear to be some outliers on the right side of the graph, with **percent_bachelors** around 85 and **no_vehilce_pct** above 30.

---

Your co-worker raises a concern that including all the numeric variables **percent_hs**, **percent_bachelors**, **disabled_pct**, and **med_hh_income** would cause collinearity issues in a GLM.

**(c)** (*2 points*) Critique the statement from your co-worker.

*Candidates struggled with this question by failing to take a clear position on the co-worker's statement or failing to provide sufficient justification for their critique. Weak responses focused only on the definition of correlation or collinearity, discussed only a subset of the variables, or failed to explain how the issue would affect GLM performance or interpretability. Full-credit responses clearly stated and supported a position, addressed all relevant variables, and explicitly connected the correlation concern to its impact on the model.*

**ANSWER:**

Your co-worker is correct about the collinearity issue for those numeric variables. For example, educational variables such as percentage with bachelor degrees would be most likely to be a subset of percentage with high school diplomas, and people with disabilities often have limited occupation options, which might result lower household income.

---

## Task 7 (*9 points*)

You are working with a coffee shop chain in Los Angeles County that plans to open stores with walk-up windows in neighborhoods with low car ownership rates. They provided the following statement:

*To increase profits, we need more high-income customers who don't own cars.*

**(a)** (*2 points*) Write down two questions to ask the company to clarify the business issue.

*Candidate performance was mixed on this task. To earn full credit, candidates need to provide two questions to help redefine the business problems. Candidates either focused on data or modeling components rather than redefining the business problem. Partial credit was awarded to candidates who gave operational type questions, such as cost of stores or drive thru windows, rather than clarifying the high-income-no-car premise.*

**ANSWER:**

*Three examples of questions are:*

i. [Question 1] What income level is considered "higher"?

ii. [Question 2] Is it acceptable to lose lower income customers to gain higher income customers?

iii. [Question 3] How would adding high-income customers increase profits? Would the company sell premium products or change its marketing?

---

You fit two Generalized Linear Models (GLMs) to predict the variable **no_vehicle**.

- Model 1: A Poisson GLM.
- Model 2: A Tweedie GLM.

The models use the following variables:

- **income_10k**: Median income in $10,000s (10k = 1)
- **disabled_pct**: Percentage of population with disabilities (10% = 10)
- **sup_dist**: Supervisory District (Reference = District 1)
- log(**no_vehicle_universe**): Log of the total number of households in the area, used as an offset in both models

Outputs for Model 1 and Model 2 are shown below:

**Model 1 (Poisson)**

```
Call:
glm(formula = no_vehicle ~ income_10k + disabled_pct + sup_dist +
    offset(log(no_vehicle_universe)), family = poisson(link = "log"),
    data = train)

Coefficients:
                     Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -2.5577550  0.0082425 -310.315  < 2e-16 ***
income_10k         -0.0826026  0.0010360  -79.735  < 2e-16 ***
disabled_pct        0.0543902  0.0003664  148.445  < 2e-16 ***
sup_distDistrict 2 -0.0063247  0.0063822   -0.991    0.322
sup_distDistrict 3 -0.0532051  0.0069108   -7.699 1.37e-14 ***
sup_distDistrict 4 -0.4204237  0.0074345  -56.550  < 2e-16 ***
sup_distDistrict 5 -0.4688121  0.0076320  -61.427  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 160394  on 1718  degrees of freedom
Residual deviance: 118630  on 1712  degrees of freedom
AIC: 128911

Number of Fisher Scoring iterations: 5
```

**Model 2 (Tweedie)**

```
Call:
cpglm(formula = no_vehicle ~ income_10k + disabled_pct + sup_dist +
    offset(log(no_vehicle_universe)), link = "log", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.4932  -1.9632  -0.5215   0.7634   8.4423
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.584242   0.078657 -32.855  < 2e-16 ***
income_10k         -0.083374   0.008409  -9.915  < 2e-16 ***
disabled_pct        0.056549   0.004333  13.050  < 2e-16 ***
sup_distDistrict 2  0.015809   0.060365   0.262    0.793
sup_distDistrict 3 -0.052494   0.063117  -0.832    0.406
sup_distDistrict 4 -0.446500   0.065143  -6.854 9.98e-12 ***
sup_distDistrict 5 -0.524101   0.067693  -7.742 1.66e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimated dispersion parameter: 4.2193
Estimated index parameter: 1.5849

Residual deviance: 8419.5  on 1712  degrees of freedom
AIC:  19163

Number of Fisher Scoring iterations:  4
```

**(b)** (*1 point*) Describe two assumptions specific to using a Poisson distribution in a GLM.

*Candidate performed well on this task. To earn full credit, candidates needed to provide the two key assumptions of Poisson distribution: the model applies to distributions with positive count variables and that the mean equals the variance.*

**ANSWER:**

- The response variable is a count variable, i.e. 0, 1, 2, …
- The conditional mean and variance are equal to each other

---

**(c)** (*2 points*) Explain why log(**no_vehicle_universe**) is included as an offset in the models rather than as a predictor variable.

*Candidate performance was mixed on this task. To earn full credit, candidates needed to explain the key technical difference that the coefficient is not fitted to the offset and provide justification on why it is preferable to use an offset in this type of situation. Many candidates provided only the general definition of an offset but failed to demonstrate the knowledge of using offsets in this case.*

**ANSWER:**

The problem focuses on the rate or proportion of households without vehicles, not just the raw count. Since census tracts have different population sizes (no_vehicle_universe), a tract with more people will naturally have more cars, even if the rate of ownership is the same. Including log(no_vehicle_universe) as an offset with a coefficient fixed at 1 effectively models the rate and allows the model to adjust for the "exposure" or size of the tract so that we can compare the density of the target market fairly across tracts of different sizes.

---

**(d)** (*1 point*) Calculate the predicted count of households with no vehicle for a district with the following characteristics using Model 2 (Tweedie):

- Income: $50,000 (**income_10k** = 5)
- Disabled Percentage: 10% (**disabled_pct** = 10)
- Location: District 3 (**sup_distDistrict 3**)
- Total Households: 1,000 (**no_vehicle_universe** = 1000)

Show your work.

*Candidate performed well on this task. To earn full credit, candidates needed to show the formula of using correct coefficients and calculate the correct predicted count. Common mistakes that received partial credit were using the log base 10 instead of natural log or ignoring the link function for the final predicted value.*

**ANSWER:**

$e\text{^}(offset + \ intercept + \sum \beta ixi)$

= $e\text{^}(\ln(1{,}000) - 2.584242 - 0.083374 * 5 + 0.056549 * 10 - 0.052494)$

= 83.07

---

Your manager asks you to build Poisson and Gaussian GLMs that include **percent_bachelors** as a predictor. The GLM specifications and performance on a test dataset are provided below:

| | |
|---|---|
| Poisson GLM | `Call:`<br>`glm(formula = no_vehicle ~ med_hh_income_10k + percent_bachelors +`<br>`  disabled_pct, family = poisson(link = "log"), data = train,`<br>`  offset = log(no_vehicle_universe))` |
| Gaussian GLM | `Call:`<br>`glm(formula = no_vehicle_pct ~ med_hh_income_10k + percent_bachelors +`<br>`  disabled_pct, family = gaussian(link = "identity"), data = train)` |

| | Residual_Deviance | AIC | MAE | RMSE |
|---|---|---|---|---|
| **Poisson_GLM** | 72441.84 | 82761.05 | 4.006 | 5.723 |
| **Gaussian_GLM** | 79683.83 | 11471.60 | 5.014 | 7.100 |

**(e)** (*2 points*) Recommend a model to the client based on performance metrics. Justify your recommendation.

*Candidate performance was mixed on this task. To earn full credit, candidates needed to consider all of the suitable performance metrics and give valid justification on the recommended model. Few to zero points were awarded to candidates who chose Gaussian due to its lower AIC without considering other metrics or alternatively chose Poisson assuming the higher AIC meant a better model.*

**ANSWER:**

I recommend a Poisson GLM.

Both MAE and RMSE are lower for the Poisson model, so it provides more accurate and more stable predictions of **no_vehicle_pct** on new data. Deviance and AIC are only directly comparable when the

models are fit to the same response variable under comparable families. Since the two models use different distribution assumptions and likelihood calculation, the residual deviance and AIC numbers are not meaningful for choosing between the two models.

---

**(f)** (*1 point*) Recommend a next step to assure the company that the model can be trusted when used on real life data.

*Candidate performed poorly on this task. To earn full credit, candidates need to reference the learning module and pick the correct step for this business problem. Many candidates provided irrelevant answers such as cross-validation, using test data, or gave a vague recommendation. Those answers received few partial points. Partial credit was awarded to answers that recalled other steps from the learning module such as adjusting the business problem, consulting with subject matter experts, gathering additional data and applying new types of models.*

**ANSWER:**

I recommend field testing the model, which is the process of implementing the model in the way it will be used but not making decisions based on it.

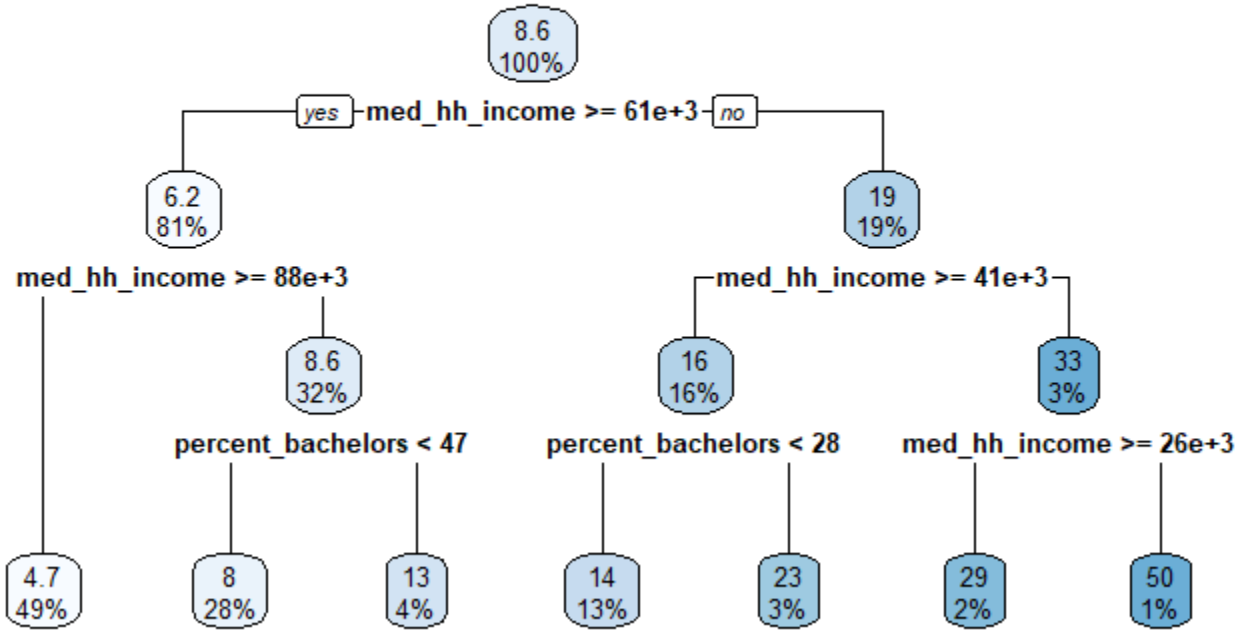
---

## Task 8 (*6 points*)

Your manager would like you to investigate whether trees can provide insight into factors related to car ownership.

You have fit a regression tree with **no_vehicle_pct** as the target variable and **med_hh_income**, **disabled_pct**, **percent_bachelors**, **percent_hs**, and **income_percap** as predictor variables. The results of the tree are depicted below:

**Exhibit 1: Initial Regression Tree**




**(a)** (*2 points*) Populate the table below with predictions from the tree:

*Most candidates performed well on this task. Some candidates provided the proportions instead of the actual prediction value and did not receive any credit. Partial credit was awarded if most cells were filled in correctly.*

**ANSWER:**

| | | **percent_bachelors** | | |
|---|---|---|---|---|
| | | **0 – 28** | **28 – 47** | **> 47** |
| **med_hh_income** | **0 – 26,000** | 50 | 50 | 50 |
| | **26,000 – 41,000** | 29 | 29 | 29 |
| | **41,000 – 61,000** | 14 | 23 | 23 |
| | **61,000 – 88,000** | 8 | 8 | 13 |

| | | | | |
|---|---|---|---|---|
| | **>88,000** | 4.7 | 4.7 | 4.7 |

---

**(b)** (*1 point*) Interpret how the impact of having a high percentage of bachelors degrees (>47%) differs between high-income and middle-income tracts.

*Candidate performance was mixed on this task. Only candidates that used the graph for interpretation were awarded full credit. The best answers cited the actual source in their interpretation.*

*Many candidates described the tree and did not provide any interpretation. No credit was awarded in these cases.*

**ANSWER:**

Suppose high income tracts to be med_hh_income > 88,000 and middle-income tracts between a med_hh_income of 61,000 and 88,000. For high incomes, there is no effect on car ownership of any underlying variable, while for medium incomes, car ownership is also split by the percentage of bachelor's degrees. Thus, we can interpret that for higher incomes, bachelor's degrees have no impact on vehicle ownership, while a higher percentage of bachelor's degrees leads to lower vehicle ownership for middle incomes.

---

To increase the number of variables included, you decide to decrease the complexity parameter (cp) and now the tree contains more splits.

```
Regression tree:
rpart(formula = no_vehicle_pct ~ ., data = train, method = "anova",
    control = rpart.control(cp = 0.005, minbucket = 10))

Variables actually used in tree construction:
[1] disabled_pct      income_percap     med_hh_income     percent_bachelors percent_hs

Root node error: 127900/1719 = 74.403

n= 1719

          CP nsplit rel error  xerror     xstd
1  0.3223761      0   1.00000 1.00120 0.085923
2  0.0958693      1   0.67762 0.70260 0.057056
3  0.0397752      2   0.58175 0.62313 0.043781
4  0.0287452      3   0.54198 0.59855 0.043631
5  0.0246791      4   0.51323 0.60094 0.044015
6  0.0105465      5   0.48856 0.56944 0.042395
7  0.0095581      6   0.47801 0.56918 0.040955
8  0.0093560      7   0.46845 0.56857 0.041441
9  0.0089391      8   0.45909 0.56738 0.041437
10 0.0075548      9   0.45016 0.54276 0.037687
11 0.0074294     10   0.44260 0.54172 0.037679
12 0.0064832     11   0.43517 0.53928 0.037632
13 0.0050000     12   0.42869 0.52380 0.036278
```

**(c)** (*2 points*) Identify the optimal number of splits according to the 1-Standard-Error Rule.

*Performance was mixed on this part. Candidates were awarded full credit only if they identified and showed the calculation to arrive at the correct model selection. Common mistakes included identifying 10 splits instead of 9 and using rel error instead of xerror. Answers were awarded partial credit provided the formula was clearly stated.*

**ANSWER:**

Threshold = min(xerror) + xstd

min(xerror) = .52380 , xstd = .036278, Threshold = .560078

number of splits = 9

---

**(d)** (*1 point*) Explain why a single decision tree may have poor model performance.

*Candidates who explained why a decision tree is prone to overfitting and identified sensitivity to training data were awarded full credit. General discussion of overfitting or bias-variance tradeoff were given minimal credit. Some candidates answered why other tree-based models were better instead of why a decision tree has poor model performance and were awarded minimal credit.*

**ANSWER:**

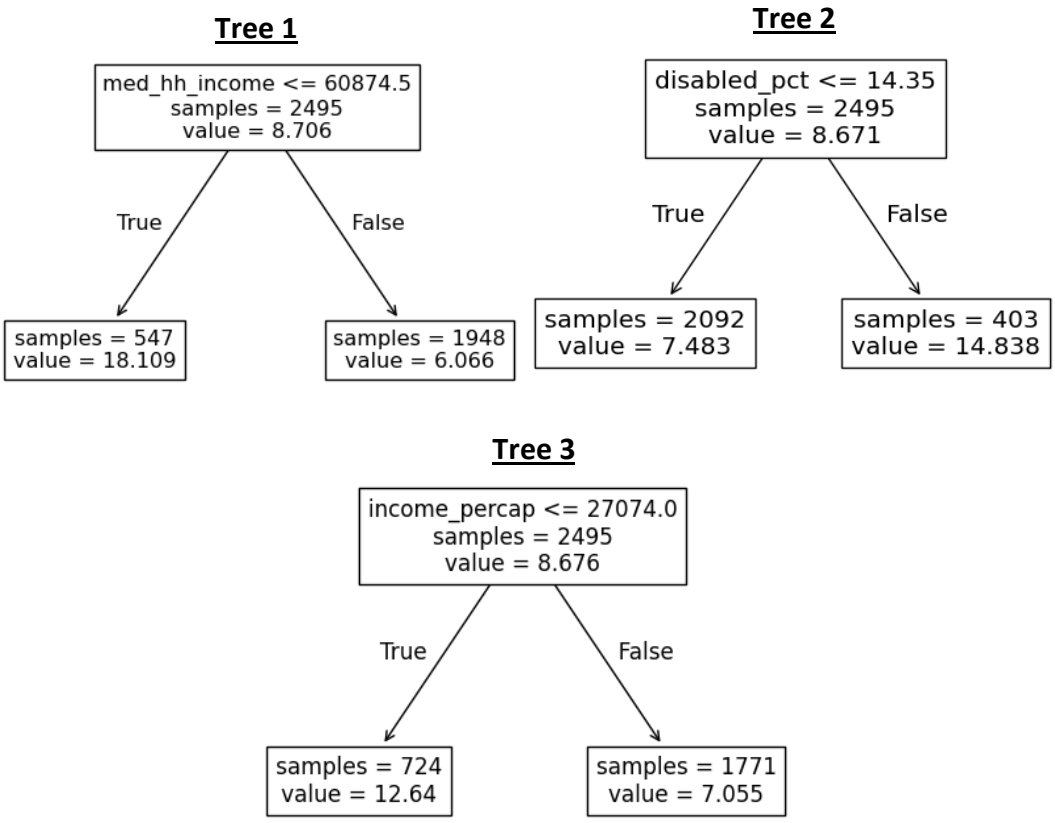
Single trees are prone to high variance where small changes in training data lead to different structures.

---

## Task 9 (*8 points*)

Your manager believes that ensemble tree models are a promising approach to modeling car ownership and asks you to build some ensemble models.

You start by building a simple a bagged tree model, which consists of the three trees below:




You are given three data points and for which trees they were used in the fitting process:

| ID | no_vehicle_pct | disabled_pct | income_percap | med_hh_income | Tree 1 | Tree 2 | Tree 3 |
|---|---|---|---|---|---|---|---|
| 1 | 2.5 | 9.5 | 33932 | 109088 | In | | |
| 2 | 15.7 | 9.4 | 27037 | 71938 | | In | In |
| 3 | 11.6 | 8.2 | 43325 | 114550 | | | In |

**(a)** (*3 points*) Calculate the Out-of-Bag MSE for the bagged model based on these three observations.

*Candidates performed very poorly on this question. Many candidates confused "in" vs "out" of bag. Some candidates got the general idea of the formula but summed and averaged in the incorrect locations.*

*Candidates should note that Excel is a good tool to use for calculation questions like this. Candidates could utilize Excel and paste a screenshot of the table as valid "work shown."*

**ANSWER:**

We start by calculating the individual out-of-bag errors:

- ID 1, Tree 2: **7.483** – 2.5 = 4.983
- ID 1, Tree 3: (7.055 – 2.5) = 4.555
- ID 2, Tree 1: (6.066 – 15.7) = -9.634
- ID 3, Tree 1: (6.066 – 11.6) = -5.534
- ID 3, Tree 2: **7.483** – 11.6 = -4.117

Then to calculate the MSE, we square and average these errors:

(**4.983**^2 + 4.555^2 + 9.634^2 + 5.534^2 + **4.117**^2)/5 = **37.193** = OOB MSE

---

You are also interested in considering other ensemble methods. Your assistant suggests boosting as an alternative to bagging.

**(b)** *(1 point)* Explain how the following hyperparameters are used in a boosted tree model

i. The number of iterations, B

ii. The shrinkage parameter, λ

*The best candidates explained the parameters in the context of the boosting algorithm. Many candidates did not provide enough context to show understanding. Vague descriptions or explaining one of two parameters correctly were awarded minimal credit.*

**ANSWER:**

Boosting is an iterative model that starts by initializing each observation (e.g. using the mean) and calculating the initial residuals. Then we fit a tree to the residuals and update the model using the new fitted model scaled by the shrinkage parameter.

B determines how many times we fit a new tree to the residuals of the previous tree. Increasing this parameter increases the predictive power of the model but also increases chances of overfitting to the training data.

λ determines how fast the boosted model learns from each iteration to reduce the residuals. As the shrinkage parameter increases, the model learns more quickly and will require fewer iterations to shrink the residuals. However, quick learning may lead to missing key relationships.

Based on client's feedback, they are interested in understanding segments where **no_vehicle_pct** is greater than 20%. You create a binary variable, **low_vehicle_own** as:

- **low_vehicle_own** = Yes if **no_vehicle_pct** > 20
- **low_vehicle_own** = No if **no_vehicle_pct** ≤ 20

Your assistant builds a classification decision tree. A confusion matrix from this model is below:

| | | Predicted | |
|---|---|---|---|
| | | No | Yes |
| Actual | No | 362 | 82 |
| | Yes | 9 | 38 |

**(c)** (*2 points*) Calculate accuracy, precision, sensitivity, and specificity from the confusion matrix. Show your work.

*Candidates performed well on this question. Candidates that provided a formula and calculated all of the metrics correctly earned full credit. Some candidates misread predict and actual values and were awarded a majority of the credit assuming the formulas were correct.*

**ANSWER:**

i. Accuracy = (TP + TN) / N = (38 + 362) / 491 = 0.8147

ii. Precision = TP / (TP + FP) = 38 / (38 + 82) = 0.3167

iii. Sensitivity = TP / (TP + FN) = 38 / (38 + 9) = 0.8085

iv. Specificity = TN / (TN + FP) = 362 / (362 + 82) = 0.8153

---

Your assistant observes that the variable **low_vehicle_own** is unbalanced between the two classes.

**(d)** (*1 point*) Explain your assistant's observation using the information provided.

*Candidates performed well on this question. Some candidates cited predicted numbers instead of actuals and were awarded partial credit. Candidates who did not provide any justification did not receive credit. Using the calculated statistics in part (c) as justification was awarded most credit. Vague answers with no supporting explanation did not receive any credit. No credit was awarded for suggesting the cutoff threshold was too low or for simply agreeing with the assistant.*

**ANSWER:**

**low_vehicle_own** being unbalanced means that there are different proportions of each level of the variable. We can see this from the confusion matrix, where there are 9 + 38 = 47 "Yes" values and 362 + 82 = 444 "No" values.

---

**(e)** (*1 point*) Describe two methods for addressing unbalanced data.

*Candidates performed well on this task. Naming methods without a description were awarded minimal credit. No credit was awarded for suggesting a change to the cutoff threshold.*

**ANSWER:**

Undersampling – Remove some of the "No" observations to balance the classes

Oversampling – Duplicate some of the "Yes" values to balance the classes

---

## Task 10 (*5 points*)

You plan to have your assistant construct a linear regression model but are concerned that due to the high variable count in the dataset, there is a risk that the model will overfit to the data. You would like to consider modifications that reduce the risk of overfitting while retaining the linear model type.

**(a)** (*2 points*) Describe two advantages of regularized regression over Best Subset Selection (BSS).

*Candidates often struggled to articulate the advantages of regularized regression, frequently rewording a previously stated benefit or simply repeating definitions of ridge regression, lasso regression, or best subset selection without explaining why the underlying properties were advantages. Common misconceptions included incorrectly stating that regularized regression binarizes variables and confusing best subset selection with iterative selection methods. Full-credit responses demonstrated an understanding of the shrinkage property of regularized regression and other key method characteristics and clearly connected those properties to their practical advantages in modeling.*

**ANSWER:**

1) Regularized regression is more computationally efficient than BSS since BSS needs to fit a model for all combinations of the predictor variables.
2) Regularization has the desirable property of shrinking the coefficients of the model, which reduces model variance and improves predictive performance.

---

You begin by investigating subsets of the predictors that could be used to calibrate a model. To begin, you ask your assistant to incrementally select counts of variables to use and calibrate a standard linear regression on each of them using all observations in the dataset. The table below summarizes the performance of each model:

| Model ID | Variables Used | RSS | AIC | BIC |
|---|---|---|---|---|
| 1 | 5 | 156305.9 | 62.8 | 63.2 |
| 2 | 10 | 140944.6 | 56.8 | 57.6 |
| 3 | 15 | 132194.6 | 53.4 | 54.7 |
| 4 | 20 | 114901.0 | 46.6 | 48.3 |
| 5 | 25 | 97713.3 | 39.9 | 42.0 |

You are concerned that the AIC and BIC continue to decrease even though they are supposed to penalize for adding additional variables to the model training.

**(b)** (*1 point*) Explain why the RSS, AIC, and BIC continue to decrease as the number of variables used increases, even though AIC and BIC penalize for additional variables.

*Candidates generally struggled with this question, and many appeared to misunderstand what the prompt was asking. Strong responses recognized that the increase in model complexity outweighed the*

*improvement in AIC and BIC. Common errors included relying on RSS as evidence of improved model performance, discussing unrelated concepts such as data leakage or overfitting to training data, and failing to address the intended tradeoff.*

**ANSWER:**

While AIC and BIC do penalize for the number of variables included in a regression, it is not always guaranteed that increasing the number of variables will lead to overfitting, especially when the number of data points is significantly larger than the number of predictors. In this case, 2495 data points is significantly larger than 5-25 predictors, so including additional predictors tends to not lead to overfitting.

---

You would like your assistant to try using Lasso regression. You think that performing a Lasso regression using an edge case for the lambda parameter could help better understand its impact.

You provide the following data points for your assistant to use:

| ID | no_vehicle_pct | Shape__Area | Shape__Length | percent_hs | percent_bachelors | disabled_pct |
|---|---|---|---|---|---|---|
| 1 | 12.9 | 12,329,806 | 14,766 | 90.1 | 27.5 | 11.4 |
| 2 | 0.7 | 29,088,657 | 30,028 | 82.1 | 33.7 | 13.1 |
| 3 | 5.4 | 7,532,651 | 12,505 | 79.7 | 31.4 | 15.7 |
| 4 | 14.8 | 3,874,399 | 8,938 | 71.7 | 23.7 | 26.7 |
| 5 | 21 | 3,090,538 | 9,751 | 78.5 | 19.1 | 18.1 |
| 6 | 2.6 | 27,694,822 | 28,701 | 91.8 | 37.2 | 16.4 |
| 7 | 14 | 67,867,979 | 41,887 | 88.5 | 25.2 | 14.0 |
| 8 | 10.2 | 12,811,195 | 16,199 | 90.9 | 29.2 | 15.2 |
| 9 | 4.7 | 179,333,612 | 27,119 | 93.1 | 44.7 | 9.6 |
| 10 | 1.3 | 5,257,057 | 9,953 | 77.8 | 20.9 | 7.5 |
| 11 | 1.7 | 139,121,772 | 63,443 | 82.8 | 37.0 | 14.7 |
| 12 | 2.4 | 21,327,765 | 24,166 | 93.7 | 35.7 | 12.1 |
| 13 | 8.5 | 13,323,876 | 16,273 | 82.0 | 17.8 | 13.2 |
| 14 | 12.5 | 21,040,409 | 18,173 | 79.2 | 26.3 | 12.1 |

**(c)** (*2 points*) Fit a Lasso regression with **no_vehicle_pct** as the target variable and a lambda approaching infinity. Provide the resulting model coefficients in the table below. Explain how you arrived at your answer.

*Candidates generally struggled with this question, with many unable to complete the calculation or incorrectly concluding that the full table could not be calculated. A common misunderstanding was recognizing that the coefficients shrink to zero while failing to correctly calculate or interpret the intercept, with some candidates incorrectly stating it was zero, unchanged, or simply the mean of the target variable without showing or performing the correct calculation. Partial credit was awarded when candidates demonstrated the correct approach but made calculation errors.*

**ANSWER:**

*Coefficients:*

| Variable | Coefficient |
|---|---|
| Intercept | 8.05 |
| Beta(Shape__Area) | 0 |
| Beta(Shape__Length) | 0 |
| Beta(percent_hs) | 0 |
| Beta(percent_bachelors) | 0 |
| Beta(disabled_pct) | 0 |

*Explanation:*

The lambda parameter controls the shrinkage of the model. As it increases, the coefficients of the resulting regression model decrease. If lambda is large enough, the coefficients can shrink to exactly 0. With an infinite lambda, all coefficients shrink to 0, leaving just the intercept term. In this case, the intercept is the average of the target variable:

(12.9 + 0.7 + 5.4 + 14.8 + 21 + 2.6 + 14 + 10.2 + 4.7 + 1.3 + 1.7 + 2.4 + 8.5 + 12.5) / 14 = 8.05

---