Project 1 GA

1. Data dictionary

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of Variable** |
| Admit | 1 = Admitted to graduate school  0 = Not admitted to graduate school | Categorical |
| GRE | GRE Score | Continuous |
| Grade point average | Grade point average | Continuous |
| Rank | Rank of an applicant's undergraduate alma mater, with 1 being the highest and 4 as the lowest | Categorical |

1. What is the outcome?

Whether the student is admitted to graduate school.

1. What are the predictors?

GRE score, GPA and rank of an applicant’s undergraduate alma mater.

1. What timeframe is this data relevant for?

Cross-sectional data

1. What is the hypothesis?

Applicant with a grade point average greater than 3.5 and alma mater undergraduate rank of 1 will get admitted into graduate school.

Exploratory Analysis Plan

1. What are the goals of the exploratory analysis?

* Check for missing data and other mistakes.
* Gain maximum insight into the data set and its underlying structure.
* Uncover a parsimonious model, one which explains the data with a minimum number of predictor variables.
* Check assumptions associated with any model fitting or hypothesis test.
* Create a list of outliers or other anomalies.
* Find parameter estimates and their associated confidence intervals or margins of error.
* Identify the most influential variables.

1. A. What are the assumptions of the distribution of data?

Normal distribution

1. How will you determine the distribution of your data?

Probability plots might be the best way to determine whether the data follows a particular distribution. If your data follow the straight line on the graph, the distribution fits your data. This process is very easy to do visually. Informally, this process is called the “fat pencil” test. If all the data points line up within the area of a fat pencil laid over the center straight line, you can conclude that your data follow the distribution.

These plots are especially useful in cases where the distribution tests are too powerful. Distribution tests are like other hypothesis tests. As the sample size increases, the statistical power of the test also increases. With very large sample sizes, the test can have so much power that trivial departures from the distribution produce statistically significant results. In these cases, your p-value will be less than the significance level even when your data follow the distribution.

1. How might outliers impact your analysis?

An outlier can affect the mean of a data set by skewing the results so that the mean is no longer representative of the data set.

1. How will you test for outliers?

Calculate first quartile (Q1), third quartile (Q3) and the in- terquartile range (IQR=Q3-Q1). CO2 emissions example: Q1=0.9, Q3=6.05, IQR=5.15. 3. Compute Q1–1.5 × IQR (=–6.825) Compute Q3+1.5 × IQR (=13.775) Anything outside this range is an outlier.

1. What is colinearity?

collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

1. How will you test for colinearity?

For quantitative variables, correlation is tested by Pearson correlation coefficient (of course, if the variable is approximately normally distributed) while for ordinal variables by Spearman rank correlation coefficient.

1. What is your exploratory analysis plan?

Tasks will include:

* Spotting mistakes and missing data;
* Mapping out the underlying structure of the data;
* Identifying the most important variables;
* Listing anomalies and outliers;
* Testing a hypothesis and checking assumptions related to a specific model;
* Establishing a parsimonious model (one that can be used to explain the data with minimal predictor variables);
* Estimating parameters and figuring out the associated confidence intervals or margins of error.

Tools and Techniques

* Clustering and dimension reduction techniques, which help you to create graphical displays of high-dimensional data containing many variables;
* Univariate visualization of each field in the raw dataset, with summary statistics;
* Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you’re looking at;
* Multivariate visualizations, for mapping and understanding interactions between different fields in the data;
* K-Means Clustering (creating “centres” for each cluster, based on the nearest mean);
* Predictive models, e.g. linear regression.

Bonus

Questions:

1. Outline your analysis method for predicting your outcome

2. Write an alternative problem statement for your dataset

3. Articulate the assumptions and risks of the alternative model