MSE 253 Statistics and Probability II — Project 2

1 Project Tasks

This project consists of three mini-projects, each focusing on a core topic in statistical analysis: hypothesis testing, experimental design, and regression modeling. You are expected to complete all three parts and submit your results as a single report (refer to instruction file). Code should be clean and well-commented, and you must include visualizations and written interpretations of your results where needed (use common sense to discern where that is).

1.1 Mini-Project 1: Hypothesis Testing [30p]

In this mini project, you will develop and test statistical hypotheses using simulated or provided data. You are encouraged to write your own hypotheses based on the questions outlined below.

1. **Proportion Test**: A survey finds that 60% of users prefer using mobile apps over desktop websites for booking appointments. Your company believes the proportion is different in your city. Use the `PRJ2_MP1_PT.csv` dataset to test whether the local preference differs from the national figure. Write your hypothesis (null and alternative), your code, and the result of your test and what it means in plain English.

2. **Difference in Means**: Two customer service strategies were implemented in your company. The customer satisfaction scores are collected and saved in `PRJ2_MP1_DM.csv`. Test whether the mean satisfaction differs significantly between the two strategies.

3. **Variance Test** You are analyzing consistency in delivery times between two delivery partners. The dataset is available in `PRJ2_MP1_VT.csv`. Test whether their delivery time variances are equal.

1.2 Mini-Project 2 [30p]

You work for a coffee company as a data scientist. The company wants to find out which roasting level leads to the highest average rating from customers. You conduct taste tests and collect scores from 50 random customers for each roast level. The results of this is collected in the `PRJ2_MP2.csv` file. Write clearly the statistical question you want to answer, your methodology in answering the question and your results. If you find that there are differences between roast levels, determine which groups differ. Finally, complete your analysis by discussing what the findings mean practically and what you suggest to your company to do next.

1.3 Mini-Project 3: Linear Regression Analysis [40p]

Download the `PRJ2_MP3_airbnb.csv` dataset either from LEARN or from Kaggle https://www.kaggle.com/datasets/aroramahima1/amsterdam-airbnb-prices-dataset/data. This dataset has 33 columns and $7,833$ rows of Airbnb listings from Amsterdam. Use this dataset to predict the price of Airbnb listings in Amsterdam using linear regression. Make sure you clean and pre-process the dataset, include exploratory data analysis, check regression assumptions, and set aside 20% of the dataset as your test data. You do not need to include all the variables (columns) in the model, but your final prediction accuracy (as long as overfitting is not an issue) will be graded. Include a correlation matrix with the model variables before you do the modeling.

Make sure you can justify any variable that is included in your model (using literature or data). Check for multi-collinearity and normality of residuals. Most importantly, interpret the regression results: Which predictors are significant? What does the coefficient of each mean in real-world terms? How good is your model? Be sure to explain your findings in simple, real-world terms. For example: "Each additional bedroom increases the price by $85 on average, holding other variables constant." Write your regression model's mathematical formula and introduce the notation in a Table.