Mark Chen

Dr. Paul Bamberg

Math 23c

11 May 2020

**Final Project Handout**

Due to social distancing, this project is done completely independently (instead of collaboratively as allowed) except for places commented specifically.

1. **The Data Set:**

The data set is acquired from a Kaggle contest, named **“**IBM HR Analytics Employee Attrition & Performance,” at the link: <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>. This data set includes a comprehensive investigation of the employment status of 1470 employees at IBM and covers information of each individual employees such as daily, monthly payment rates, year of employment, job involvement and satisfaction levels, job categories and so on (a total of 34 columns). Overall, the 34 columns of 1470 samples form a desired data frame meeting the requirements listed for the assignment.

1. **Summary of Analyses Done:**

Using some of the statistical tests I learned in class, I did mainly four rounds of analyses to explore some of the problems the society heatedly debates, such as the relationship between gender and salary and between years of experience and salary. Admittedly, since the data is collected inside of IBM, so the ranges of industries and companies are both limited. I will discuss two main topics.

* 1. Gender vs. Hourly Rate
     1. I ran a permutation test that tests the relationship between gender and the hourly payment (I chose hourly payment to minimize external influences to just focus on the payment bracket)
     2. P-Value = 0.4928, marking the lack of sufficient evidence to reject the null hypothesis that hourly payments are not based on gender.
  2. Total Years of Working vs Monthly Income
     1. I selected Monthly Income for this category because total years of working should also influence the overall mortgage levels and so on that other types of salary categories, such as hourly rate, cannot reflect.
     2. I used both a Student T-Test and a Permutation Test as a comparison between a traditional statistical test and a simulation test. Both results show statistics that rejects the null hypothesis that these two columns of information are not related.
     3. P-Value = 0; the confidence interval of monthly income is (6250.882, 6732.421); T-Statistics = 52.86653.