**MGMT 58600 Python Programming**

**Lab 6**

**Lab Purpose:**

The purpose of this lab is to provide experience with Pandas and Python notebooks.

**Lab Instructions:**

*Description:*

As part of its acquisition and expansion strategy, KSM is investigating acquiring an auto insurance company that will allow them to complement their auto sales by also providing insurance to their customers (and non-customers). You’ve been provided a sampling of insurance claims that have been made by the target company’s customers. This will allow you to do some initial investigations (exploratory data analysis … EDA) to answer some preliminary questions that will be the basis of more detailed analysis.

Using Jupyter notebook, and borrowing on your newfound knowledge of Pandas, create the dataframe(s) necessary to answer several preliminary questions and investigate whether there’s an possible merit to several “premises” or “conventional wisdom” claims regarding drivers, automobiles, and their insurance/driving behavior.

About the dataset:

* Gleaned from a data subset of claims made by customers in several states that are currently serviced by the target company (the company that KSM may acquire to facilitate a quick expansion into the automobile insurance field). While the states represented are only Oklahoma, Kansas, Nebraska, Iowa, and Missouri, they are representative of the data contained in the whole company claims database.
* For clients with unemployed as their status, the income field has been left blank. You will need to compensate for this as you process the data. To help you with this task, the target company has indicated that “unemployed” only reflects the current state of the insured at the time that the policy is written and is actually more akin to a person who is changing jobs on hiatus or temporary company furlough with indication of near-term rehiring. They have confirmed that the typical income (when the unemployed client is actually employed) is comparable to that of employees that have “employed” status (note that there are other categories as well including disabled, medical leave, and retired and these ARE NOT necessarily in line with unemployed figures).
* The data is stored in an Excel spreadsheet called “insurance.xlxs.”

About your program:

* Must be developed/save in Jupyter notebook. Note that, once you have finalized working in Jupyter notebook, you’ll need to then upload it to the Brightspace assignment submission area.
* Must create dataframes for the process of analysis.
* Must use the markdown capabilities in Jupyter notebook to comment your code as well as present your findings. Findings should be clearly identified in red (markdown findings in red would have <font color = ‘red’> in front of the finding and </font> to conclude the finding. This will cause everything between to be highlighted in red. The markdowns should be properly placed right before the actual code segment that carries out a specific activity or addresses a particular question/premise.
* Must adequately account for missing data. This can be done using a couple of different ways; however, you may not 1) drop records with missing data or 2) assign zero (or some other arbitrary number) and it must be accounted for within the notebook (i.e., you can’t pull out your calculator, figure out a number and hardcode it into a variable). Be sure to state HOW you are accounting for the missing values in your markdown.
* Must provide general stats about continuous variables including minimum value, maximum value, standard deviation, mean, count, and the 25th, 50th, and 75th percentiles. This should be shown both BEFORE and AFTER you have accounted for the missing income data.
* Must use at least one plot or graph to substantiate your findings in an aspect of the EDA.
* Must address the following premises or “conventional wisdoms” regarding drivers, automobiles, and their insurance/driving behavior:
  + ~~Mother Nature trumps human mess-ups. What does the data say about the average claim reason of hail vs. the other types of reasons? This is based on claim amount (not total claim amount).~~
  + ~~People who buy more expensive policies tend to have higher cost claims (note that basic is the cheapest policy, extended is mid-range, and premium is the most expensive policy type). This is based on claim amount (not total claim amount).~~
  + ~~Married folks are safer drivers than single folks (look at the counts as a percentage of collision claims only to address this premise).~~
  + ~~City drivers are more reckless (ironically they get in more wrecks) than suburban or rural drivers (look at the counts as a percentage of collision claims only to address this premise).~~
  + ~~Big cars equal big claims. What does the data say about the vehicle size of large vs. the other vehicle sizes? This is based on claim amount (not total claim amount).~~
  + ~~What does the data say about education and the claim amount? This is based on claim amount (not total claim amount).~~
  + ~~Of the states represented in our sample, is there one that we should probably look at dropping coverage of its drivers? This is based on claim amount (not total claim amount).~~

**Grading Criteria**

|  |  |
| --- | --- |
| **Element** | **Points** |
| Link to insurance.xlsx correctly established | 5 |
| Developed properly in Jupyter notebook | 10 |
| Correct use of dataframes | 5 |
| Adequately accounts for missing data | 10 |
| General stats about continuous variables | 5 |
| At least one plot or graph supporting a finding | 10 |
| Premises: |  |
| * Mother Nature Premise | 5 |
| * Expensive Policies Premise | 5 |
| * Married Folks Premise | 5 |
| * City Drivers Premise | 5 |
| * Big Cars Premise | 5 |
| * Education Findings | 5 |
| * State Drop Suggestion | 5 |
| Program executes correctly and effectively | 10 |
| Comments/Markdowns | 10 |
| Total Points Possible | 100 |