UDACITY DATA ANALYST NANODEGREE

Data Visualization Project

By:

Janely Padillo

1. **INTRODUCTION**
    a. Important information about the dataset was introduced, which included the dataset content, structure as well as the source.
    b. Objectives during the analysis as well as determining the independent variable I am interested are also pinpointed during this stage.
    c. A table of content was created to outline what will be explored in this project.

2. **GATHER/READ DATA**
    a. Import libraries necessary for data exploration, which includes Numpy, Pandas, Seaborn and Matplotlib.
    b. Read-in csv file from Prosper Marketplace Inc. into a dataframe using pd.read_csv() method.

3. **ASSESS**
    a. Visual assessment was performed for preliminary exploration using .head(), tail() and sample() methods.
    b. Programmatic assessment was performed using .info() and .describe() methods.

4. **WRANGLE**
    a. A copy of the dataframe was created in preparation of data wrangling to be performed.
    b. Data cleaning was performed as follows:
        i. Created a new column that is the mean of Credit Score Range Upper & Credit Score Range Lower using a mask. Deleted the two columns using .drop().
        ii. Combined ‘Others’ and ‘Professionals’ values under Occupation into ‘Other Professionals’ using a mask.
        iii. Updated ‘Not Displayed’ to ‘Not Reported’ using a mask.
        iv. Dropped Credit Grade value ‘HR’ even though I ended up not using the data.

v. Created a copy of the dataframe for more data wrangling after univariate exploration using .copy().

vi. Converted the credit score from float to integer using .astype()

vii. Converted the stated monthly income from float to integer using .astype()

viii. Removed outlier from stated monthly income.

ix. Removed ‘Other Professionals’ under occupation since it is an outlier and does not provide any additional information.

5. **EXPLORE**

   a. Univariate exploration:

      i. Used seaborn’s barplot to check for missing values and determine whether it warrants action. None of the variables I am interested in analyzing is missing significant values so no action is necessary.

      ii. Used seaborn’s countplot for credit scores, which appears to have a normal distribution. The most frequently occurring credit score is 699.

      iii. Used .value_counts() and indexing to only select the top 10 most common occupations. Afterwards, I created a countplot using seaborn for the occupations. The value with the highest frequency is the ‘Other Professionals’. I make a note to drop ‘Other Professionals’ from the occupations in order to ensure that the plot is less skewed.

      iv. Used matplotlib to create a histogram of Debt-to-Income ratio, which is highly skewed. Majority of the values are found between 0.1 and 0.5.

      v. Also used matplotlib’s histogram for monthly loan payments, which, similarly enough, is right skewed.

      vi. Used matplotlib’s pie to visualize the income range proportion, which I indexed to only show the top five. The pie chart shows that majority have an income between 25,000-49,999.

      vii. Used matplotlib’s histogram to visualize the stated monthly income, which, like its predecessors, is also skewed right.

   b. Bivariate Exploration

      i. Used seaborn’s violinplot() to visually explore the relationship between occupations and credit scores. With the exception of the value ‘Other Professionals’, the rest of the occupations are normal in distribution. Majority of the credit scores hover between 650-750.

      ii. Used seaborn’s regplot() to explore the relationship between credit score and debt-to-income ratio. The regression plot shows a slightly negative linear relationship between the two.

iii. Used matplotlib’s hist2d() to create a heat map in order to assess any correlation between credit score & monthly loan payment, which seems to show a positive relationship.

iv. Used seaborn’s regplot() to visualize the relationship between stated monthly income and credit score. It shows a positive linear relationship between the two.

v. Created a function called boxgrid to create a series of boxplots within a PairGrid in order to compare the relationships between stated monthly income, monthly loan payment, debt to income ratio and occupation. In this visualization, we can see that there are significant outliers in all numeric data in relation to the categorical variable ‘Occupation’. For both stated monthly income and monthly loan payments, borrowers who are executives rank the highest but the debt to income ratio seems to be similar across the board.

c. <u>Multivariate Exploration</u>

i. Used .isin() to divide top 10 occupations into two separate, custom dataframes so as to avoid overplotting. Used seaborn’s regplot to determine the relationships between credit scores, debt to income ratio and occupations. Interestingly enough, the types of occupation seem to have an effect on whether or not the relationship between the two numeric variables are positive or negative. For those who work as administrative assistants, skilled laborers, salespeople(retail), retail managers and clerks, it seems the relationship is positive whereas accountants/CPAs, executives, computer programmers, teachers, analysts and salespeople (commmission) trend towards negative.

ii. Used .FacetGrid() to explore the relationship between stated monthly income and monthly loan payments in relationship to credit scores. As the visualization shows, the relationship between stated monthly income and loan payments is positive for all credit scores.

iii. Used PairGrid() to visualize the relationship between credit score, monthly loan payments and debt-to-income ratio. It appears that credit score shows a strong positive relationship with monthly loan payments but a weak, negative relationship with debt-to-income ratio. On the other hand, monthly loan payments appear to have a positive relationship with debt to income ratio.