**Data Description**

**Naixin Zhang**

The western New York area hospital Discharge dataset from 2014 to 2016 is used for analysis and building the models in this project. The dataset is publicly available from the website of HealthData(https://healthdata.gov). This dataset contains record-level details regarding the discharge of inpatients in the western New York area including information such as age, gender, race, facility ID, diagnosis, length of stay, payment typology and son on. Table 1 provides a brief introduction of the features' in the dataset.

**Table1 data's feature introduction**

|  |  |  |  |
| --- | --- | --- | --- |
| features | description | unique values | variable type |
| Health Service Area | A single county or group where the service happened | 'Western NY' | / |
| Hospital County | The county where the hospital is located. | 'Allegany', 'Cattaraugus', 'Chautauqua', 'Erie' | / |
| Operating Certificate Number | All hospitals in New York State need Certificates of Need and Operating Certificates to be open | 226700,  228000  …… | / |
| Facility ID | The unique number which is assigned to each hospital in the health  county. | 218.,  216.,  …… | / |
| Facility Name | The name of the hospital | / | / |
| Age Group | the age of the patient | '70 or Older',  '50 to 69',  '30 to 49',  '18 to 29',  '0 to 17' | category |
| Zip Code - 3 digits | The zip codes in each health county | / | / |
| Gender | The gender of the patient | 'F', 'M', 'U' | category |
| Race | The race of the patient | 'White', 'Black/African American', 'Other Race' | category |
| Ethnicity | The Patient’s ethnic background | 'NotSpan/Hispanic', 'Spanish/Hispanic',  'Unknown' | category |
| Length of Stay | The Patient’s stay at hospital | Range from 1 to 120+ | continuous |
| Type of Admission | The reason for the patient’s admission | 'Elective',  'Urgent', 'Emergency', 'Newborn',  'Not Available' | category |
| Patient Disposition | Where patients should be sent after the treatment | Home or Self Care', 'Short-term Hospital', 'Skilled Nursing Home',  ……  'Critical Access Hospital',  'Another Type Not Listed' | / |
| Discharge Year | The year the patient was discharged from the hospital. | '2014',  '2015',  '2016' | category |
| CCS Diagnosis Code | a tool for grouping patient diagnoses into a manageable number of clinically meaningful categories | / | / |
| CCS Diagnosis Description | / | / | / |
| CCS Procedure Code | The Clinical Classifications Software (CCS) is a tool for  grouping patient procedures into a manageable number of  clinically meaningful categories | / | / |
| CCS Procedure Description | / | / | / |
| APR DRG Code | An abbreviation for all patients Refined Diagnosis Related  Groups (DRG), which assigns specific code for each diagnosis  group. | / | / |
| APR DRG Description | / | / | / |
| APR MDC Code | An abbreviation for all patients Refined  Diagnosis Related Groups (DRG), which assigns specific code for each diagnosis group. | / | / |
| APR MDC Description | / | / | / |
| APR Severity of Illness Code | Severity of Illness | 1, 2, 3, 4 | Ordinal |
| APR Severity of Illness Description | Illness Description | 'Moderate', 'Minor', 'Major', 'Extreme' | Ordinal |
| APR Risk of Mortality | Risk of Mortality | 'Minor', 'Major', 'Moderate', 'Extreme' | Ordinal |
| APR Medical Surgical Description | Medical Surgical Description | 'Medical', 'Surgical', 'Not Applicable' | category |
| Payment Typology 1 | It is used to identify the payer expected to pay the major percentage of the patient's bill | 'Self-Pay', 'Blue Cross/Blue Shield', 'Medicare', 'Medicaid',  'Federal/State/Local/VA', 'Private Health Insurance',  'Miscellaneous/Other', 'Department of Corrections',  'Managed Care', 'Unspecified' | category |
| Payment Typology 2 | Allows more particularity for the source of payment  classification. | same as above | category |
| Payment Typology 3 | Allows classification at highest level of specificity | same as above | category |
| Attending Provider License Number | Used to recognize the physician or other health care professional mainly responsible for the care of the patient | / | / |
| Operating Provider License Number | For operation legally | / | / |
| Other Provider License Number | / | / | / |
| Birth Weight | Birth Weight | 0-9900 | category |
| Abortion Edit Indicator | Whether the patient has an abortion or not. | 'N'  'Y' | category |
| Emergency Department Indicator | / | / | / |
| Total Charges | The initial, individual list prices | / | continuous |
| Total Costs | Expenses incurred by a hospital in providing patient care. | / | continuous |

As we can see from Table 1, there are many features in the dataset, and most of them are category variables. To build good intuition about our dataset, in this section, we will investigate two aspects: data preprocessing and data visualization.

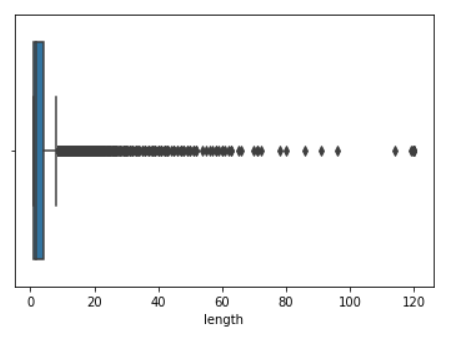
In data preprocessing, four steps were performed: (1) Extract the data we need (2) Checking for missing values (3)Checking for outliers (4)Checking for imbalance data (5) Conducting a correlation test to eliminate the redundant features. Details are as follows.

Firstly, we merged the three years of data into a new data frame using pandas in python. Now we have 7057910 rows and 37 columns. Based on our project objective, we'd like to investigate if health insurance status influences the hospital length of stay for patients with asthma conditions. Then, we extract patients who have asthma conditions. We got 885925 inpatients and 37 columns after extraction. Since there are many useless features like code or description for our objective, we would like to drop some of them. And We only have 22 columns now.

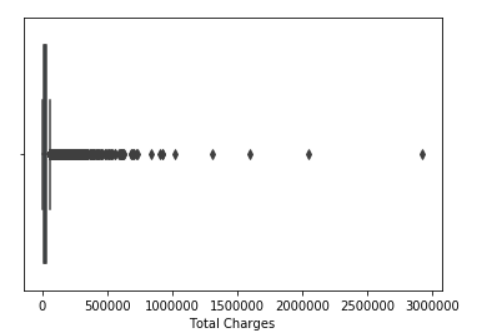
Secondly, we will check if there are missing values. Since only one column has a small number of missing values, so we drop it. Then we further detect that in the " Birth Weight" column, there are 3200 rows are 0, which is weird. So, we will drop this column.

Thirdly, since there are just three columns are numerous variables: "length of stay", '' Total Charges "and "Total Costs" and We notice that the length of stay column is string type and the days above 120 displays "120+ ", we need to transfer it to integer type firstly. Then, as we all know, if there is an outlier, it will be plotted as a point in the boxplot, but other populations will be grouped and display as boxes. Figure 1 to Figure 3 shows the distribution of the three features. For the "length of stay" variable, most of them are 0-5 days which are under the Q1 percentile. It indicates that we can't simply drop them. So as "Total Charges" and "Total Costs". Moreover, we will assume that there are no outliers. After pre-process we have the dataset like Figure 4 below.

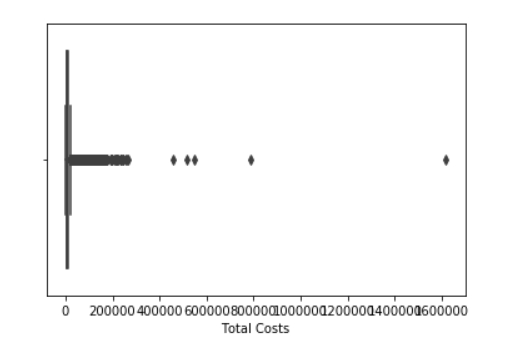
**Figure1 Distribution of the length of stay for patients**



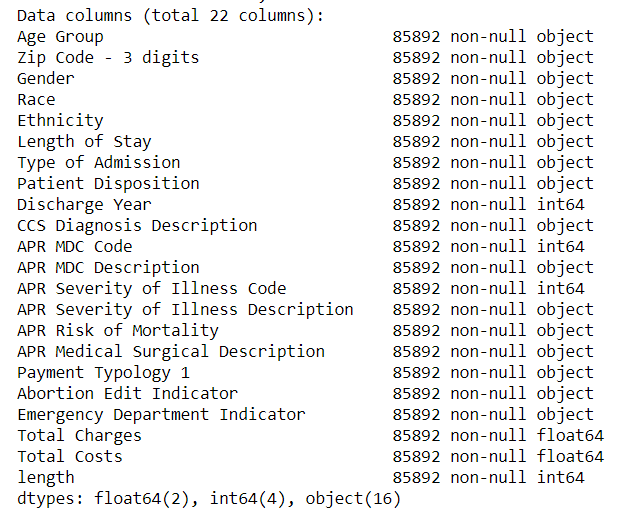
**Figure2 Distribution of the total charges for patients**



**Figure3 Distribution of the total costs for patients**



**Figure 4 After the first three steps the data we have**

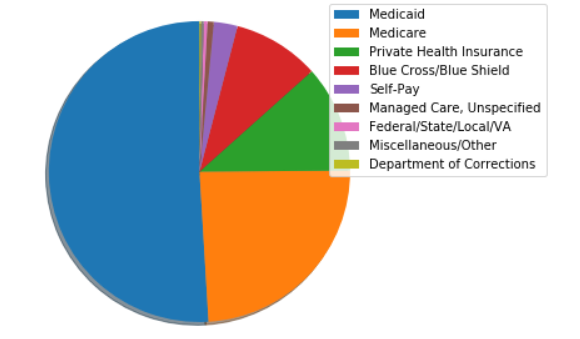


Fourthly, since we are interested in the relationship between the length of stay in hospital and payment typology. We can detect if there are any imbalance issues by checking the distribution of the payment method. We have ten different payment method indicated in Table 2 and Figure 5.

**Table 2 The count of each payment method**

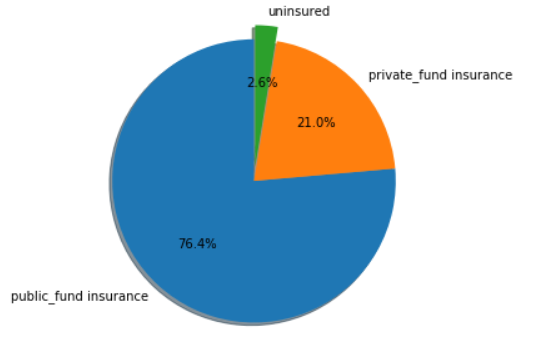
|  |  |
| --- | --- |
| **Payment Method** | **count** |
| Medicaid | 43690 |
| Medicare | 20697 |
| Private Health Insurance | 9871 |
| Blue Cross/Blue Shield | 7983 |
| Self-pay | 2182 |
| Managed Care, Unspecified | 577 |
| Federal/State/Local/VA | 344 |
| Miscellaneous/Other | 253 |
| Unknown | 165 |
| Department of Corrections | 130 |

**Figure 5 The percentage of each payment method**



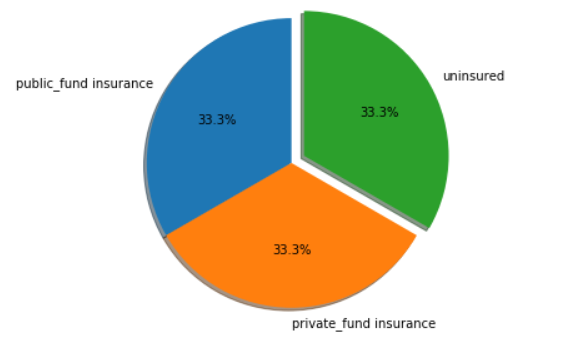
Because there are too many payment methods, it looks a bit confusing. We summarize the ten-payment method into three categories: public insurance payment, private insurance payment and uninsured(self-payment), we can see from figure 6.

**Figure 6 The bar chart of the payment method**



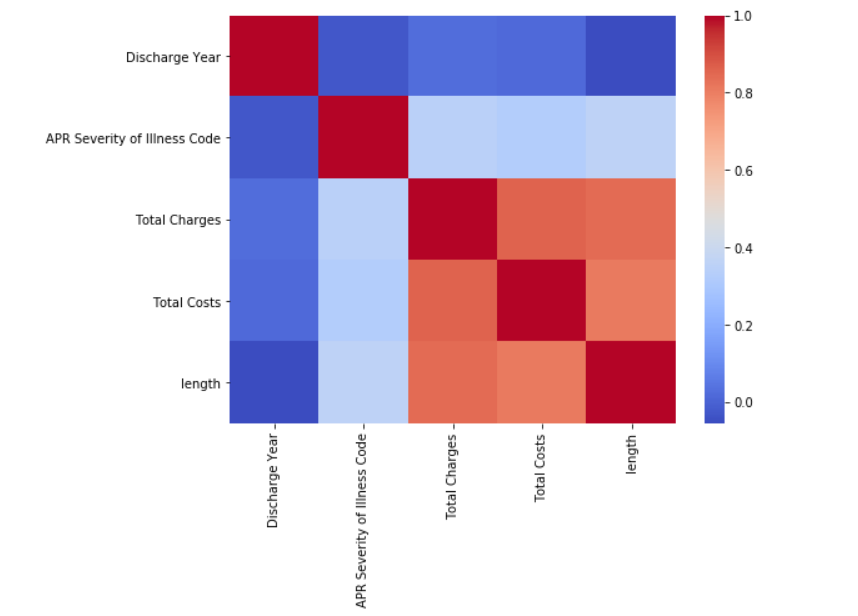
Since the distribution is highly imbalanced, half most of them are paid from public fund insurance. I use the resampling method by only randomly selecting the same number as uninsured inpatients from private fund insurance and public fund insurance patients based on each year as the research objects. Now the data is balanced. We can check it from Figure 7.

**Figure7 The distribution of each payment method**



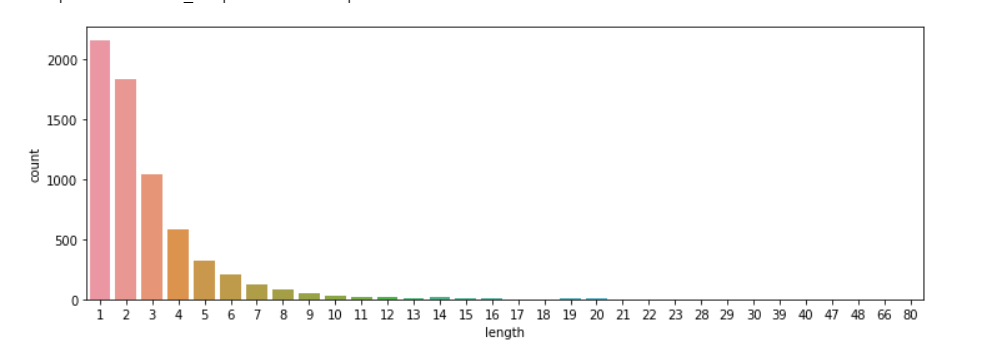
Lastly, Multicollinearity occurs when independent variables in a regression model are correlated. We will use a heatmap in Figure8 to detect if there are correlations among every two variables. The figure indicates that the total charges and total costs are highly correlated, and the length of stay in the hospital is also highly correlated with total charges and total costs. So in the regression part, we need to be careful about the correlation among them.

**Figure 8 The correlation between every two variables**



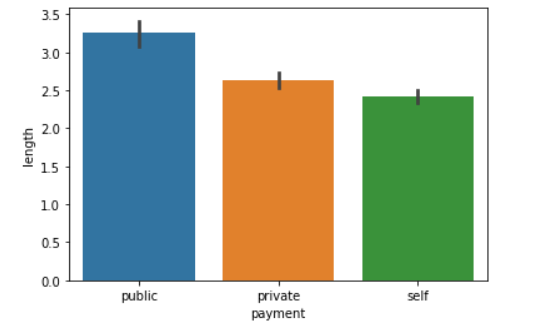
As we have finished the data pre-processing part, before doing regression analysis, let's further doing some data visualization. This will make big data easier for our brain to understand, and to detect patterns, trends.

**Figure 9 The distribution of length of stay in Hospital**



As we can see from Figure9, most of the patients stay in hospital within one week.

**Figure10** **The relationship between length of stay in hospital and payment method**



**Figure11 The count plot of the length of stay in hospital based on payment method**

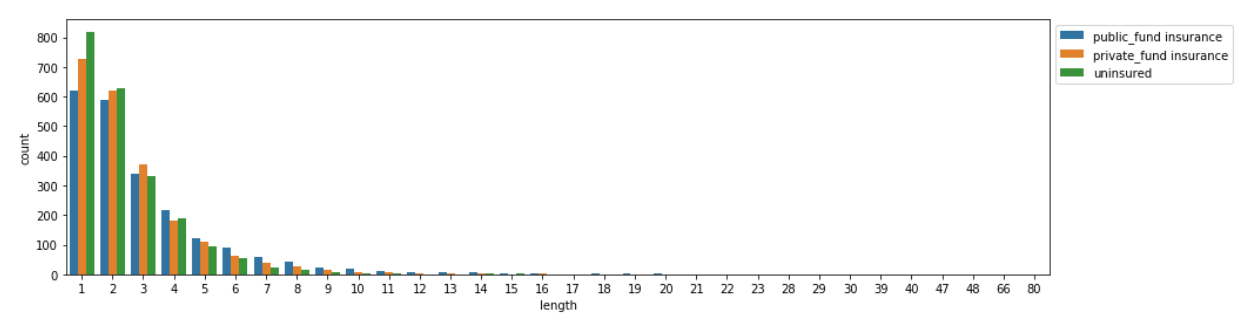


Figure10 indicates the relationship between the length of stay in hospital and payment. It seems public fund insurance paid patients tends to stay in the hospital longer than the other two payment method. Uninsured patients tend to stay in hospital in a shorter time. Furthermore, Figure11 tells us more detail information. The uninsured patients tend to discharge on the same day after admission.

**Figure12 The number of race and age group**

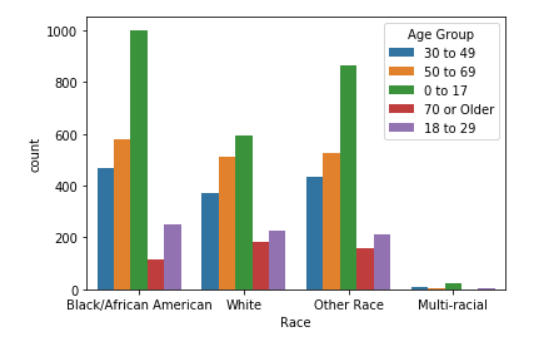


Figure12 indicates teenager accounts for a large proportion in each race group in our dataset. 70 and older patients account for a small proportion.

**Figure13 The relationship between length of stay in hospital and age group**

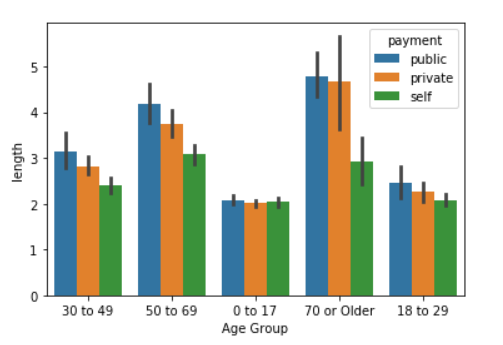


Figure 13 indicates the elderly patients tend to stay longer in hospital and young patients tend to stay longer in hospital. However, in elderly patients, there are less uninsured patients compared with other age groups; these patients tend to discharge within three days. It shows that age does influence the length of stay in the hospital.

**Figure14 The relationship** **between length of stay in hospital and Race**

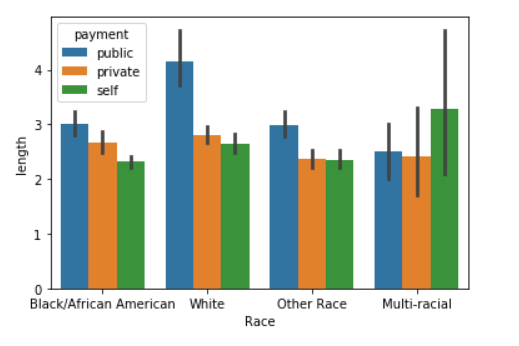


Figure14 indicates the relationship between the length of stay in hospital and race. White patients have the highest proportion to have public insurance which might make them stay in hospital longer compared with another racial group. Moreover, multi-racial patients have a high proportion of no medical insurance, but they stay in the hospital longer.

**Figure15 The relationship between total charges and Gender**

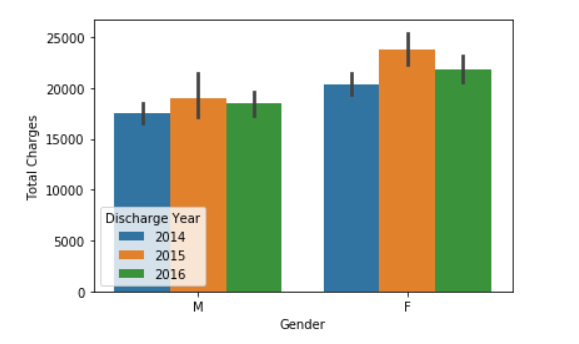
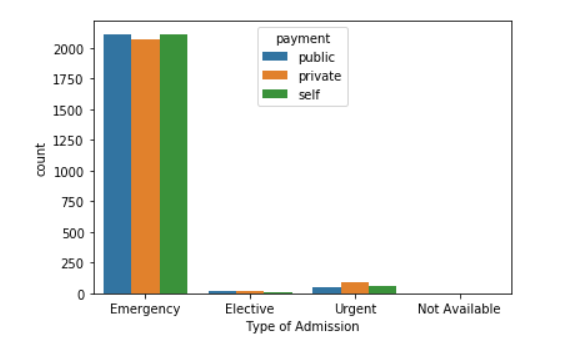


Figure15 indicates that female tend to be charged more compared with male, and during the three years, 2015 has the highest total charges for both male and female.

**Figure16 Count plot of the type of Admission**



Generally, there are three major types of hospital admissions, emergent and elective. Emergent usually happens when a patient seen in the emergency department is subsequently admitted to the hospital. Elective hospital admissions occur when a doctor requests a bed be reserved for a patient on a specific day. The patient then checks in at the admissions office and does not go to the emergency department. For urgent admission patients, they may arrive at the hospital by their transport or in an ambulance. This is known as an ‘unplanned presentation’. Figure 16 indicates that most of the patients are emergency type, there is no significant difference among each type.

**Figure17 Relationship between length of stay in hospital and type of admission**

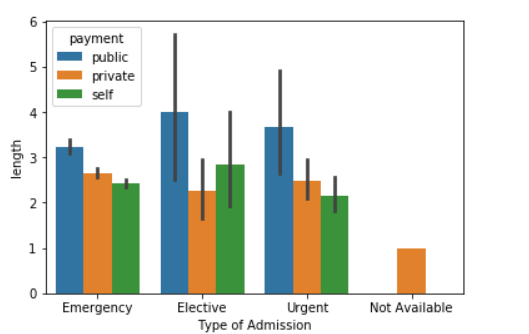
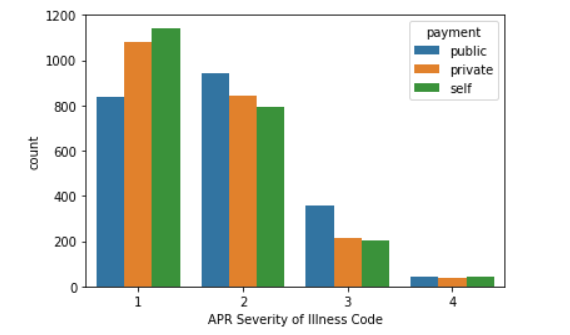


Figure 17 gives more detail information about the length of stay in hospital and type of admission. In general, public fund insurance patients tend to have longer time stay in hospital compared with private fund insurance patients and uninsured patients. However, Among the elective type admission patients, uninsured patients stay in hospitals longer than private fund insured. It is a little weird.

**Figure18 The number of patients at each severity of illness**



As mentioned in Table 1, the severity of illness code is defined as the extent of organ system derangement or physiologic decompensation for a patient. It gives a medical classification into minor, moderate, major, and extreme. The larger the code, the more severe of the illness. Figure17 indicates the most patients have minor or moderate status. Strangely, most uninsured patients have minor or moderate status. Since public fund insurance patients tend to stay in the hospital longer, we wonder if the private fund patients have a pre-selection of the patients according to their health status. Figure 18 does show that patients with private insurance tend to have less severity of the illness.

**Figure19 The relationship between length of stay in hospital and severity of illness**

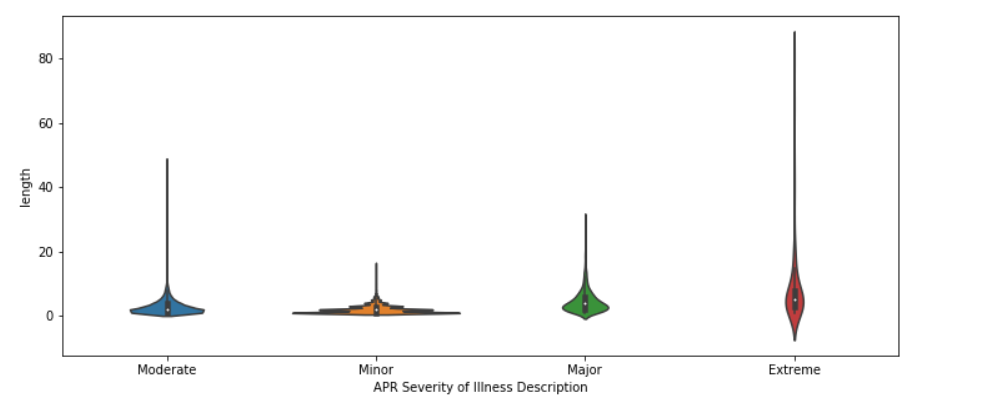


Figure19 indicates the patients who have more severe illness tend to stay in the hospital longer. Moreover, it has the most significant variance. Patients with minor status have wider sections of the violin plot, represent a higher probability that members of the population will take on the given value.

**Figure 20 The relationship between length of stay in hospital and treatment method**

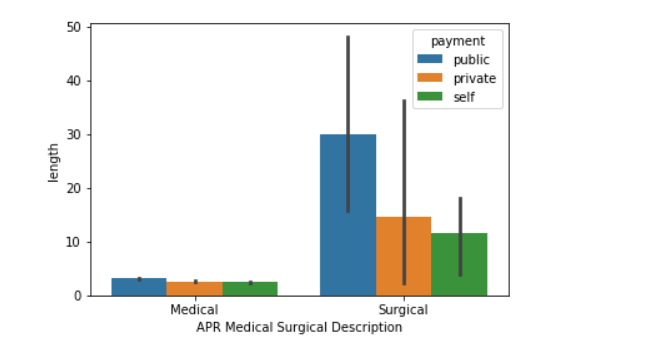
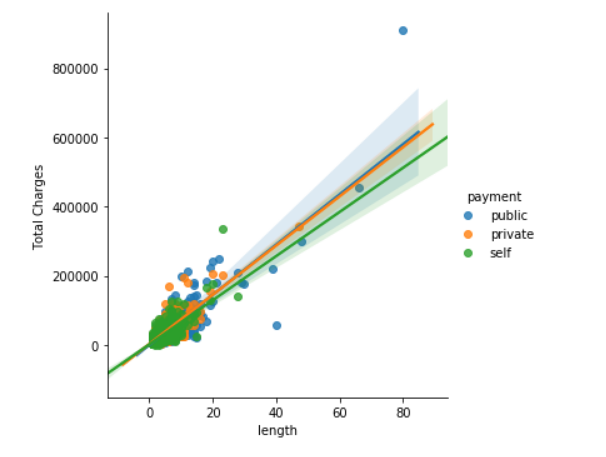


Figure 20 indicates the most of the patients are treated as the surgical method. For patients treated as the medical method, there is no statistical difference in the length of stay in hospital among different payment method. However, for patients treated as the surgical method, most of them paid by public fund insurance.

**Figure 21 The relationship between total charges and length of stay in hospital**



**Figure 22 The relationship between total charges and length of stay in hospital based on different payment method**

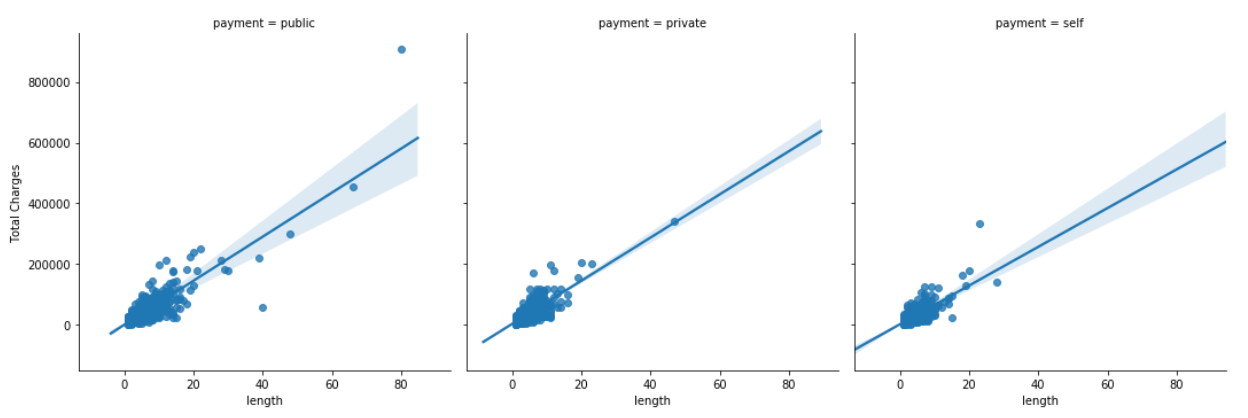


Figure21 indicates that the positive relationship between total charges and length of stay in the hospital, patients who stay longer in the hospital tends to charge much. Figure 22 shows that there is no significant difference among each payment method.

**Figure 23 The relationship between payment and length of stay based on gender**

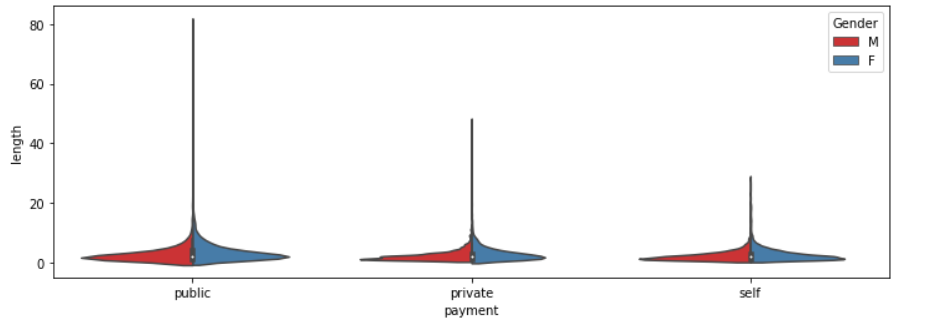


Figure23 indicates there is no evidence shows male and female has a difference in the length of stay in hospital.

General comments:

Nice job! You provided a lot of details about your data, and in a thoughtful way. The more you understand your data, the more likely you may generate useful analysis. It is always a good idea to start with more details, however, at certain point, you need to start trimming down unnecessary details. So keep in mind when you move along about what is the main story of your analysis, and prioritize the details to be reported in the final poject.