CSP 554 Big Data Technologies

Project Report

On

Visualization of Healthcare Analytics

By

Name of Student	CWID
1. Akash Didigi Kashinath	A20524076
2. Patel Henilkumar Hareshbhai	A20513297
3. Andrews Acheampong	A20516669

Under The Supervision
of
Prof. Joseph Rosen

College Of Computing

Illinois Institute of Technology
Chicago, Illinois.

December 2022

CONTENT

Section I: Project Details

1.1- Project Topic: Apply a range of big data tools to explore some interesting data sets and derive insights from them. Ingest data, apply transformations, profile the data, summarize it, visualize it.

1.2- Application Subject Area: Healthcare Management

Problem : In this era, the health is considered as a predominant part of every one's life. During the covid-19 situation, the health care industries play vital role to prevent the covid-19 infection, but from the case study, some of the hospitals and healthcare industries have suffered from the managing the patient, because of the less availability of the sources. In this project, we analyze the patient condition based on the patient illness, length of stay in the hospital, so we can arrange the better treatment for them to avoid the risk of death.

Solution : The patient has high risk, at the time of admit, it will get the higher attention of staff, to reduce the infection, also hospital staff can improve the treatment plan for reducing the chance of death.

1.3- Data Set Source : Healthcare Analytics-2 Data taken from the Kaggle

- This data set has more than 3 lakhs data rows, and this data is collected during the covid-19 pandemic situation, in this data set there are different features like case_id, hospital_code, hospital_type_code, city_code_hospital, hospital_region_code, Available Extra rooms in hospital, Depatment, ward_type, ward_facility_code, Bed grade, Patient id, Type of admission, Severity of illness, Visitors with patient, age, admission deposit, and stay. [2]
- Now based on the patient condition(taking stay as data), and Severity of illness, we can arrange the better treatment for the patient to avoid the risk of death.
- Other features like hospital_code, city_code_hospital, hospital_region_code, and so on features are helpful for finding best hospital and treatment in the same area.
- In our project, the column Stay is the target variable, because here we are trying to predict a time, during that time the new patient comes.

1.4- Analysis at First Glance: In Primary case, I found the two columns(Bed Grade, City_Code_Patient) having null values, and missing ness is the MNCR type values.

- From Further analysis of data, I concluded that, there is no more than 2% null values are present in the datasets.[8]
- To make the project more interactive and easier, we used the not only plotly but also altplot, because these are very helpful for the detail observation of the data.
- Now for the visualize the categorical columns, there is a hiplot library available, so we can implement it on our project.
- Now if we want to remove the outliers of the data, and normalization of the data, we can use the parelly plot.
- In my opinion, from looking at data, in this time, to reduce the spread of any disease among the other people, maintaining the proper healthcare system, and the number of machines require for to combat the disease needs to be available, in the hospital management.

1.5- Features of Datasets :

Attributes	Description	Data Type
Case-id	Case id registered in hospital	INT
Hospital_code	Unique code for the hospital	INT
Hospital_type_code	Unique code for the type of Hospital	INT
City_Code_Hospital	City Code of the Hospital	INT
Hospital_region_code	Region Code of the Hospital	STRING
Available Extra Rooms in Hospital	Number of Extra rooms available in the Hospital	INT
Department	Department overlooking the case	
Ward_Type	Code for the Ward type	STRING
Ward_Facility_Code	Code for the Ward Facility	STRING
Bed Grade	Condition of Bed in the Ward	INT
Patientid	Unique Patient Id	INT
City_Code_Patient	City Code for the patient	INT
Type of Admission	Admission Type registered by the Hospital	STRING

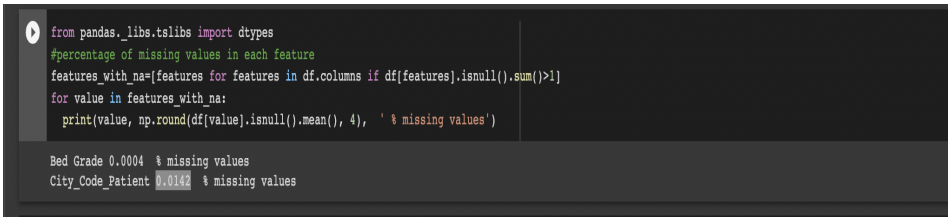
Severity of Illness	Severity of the illness recorded at the time of admission	STRING
Visitors with Patient	Number of Visitors with the patient	INT
Age	Age of the patient	INT
Admission_Deposit	Deposit at the Admission Time	INT
Stay	Stay Days by the patient	INT

Section II: Proposed Approach:

1. **Data Source:** There are thousands of datasets available on the online platform. Still, we are looking for the datasets whose insight helps predict the future, so after going through the Kaggle, we found the healthcare analytics datasets which contain the recent covid-19 patient details; so, by finding insight from this data, if this kind of situation will happen in future, we can control it.[2]

2. **Data Cleaning:** Data cleaning is the process in which we find duplicate values, missing values, mismatched data format values, and many more; after finding these values, we substitute those values by using any one technique, like the mean, maximum, and average.

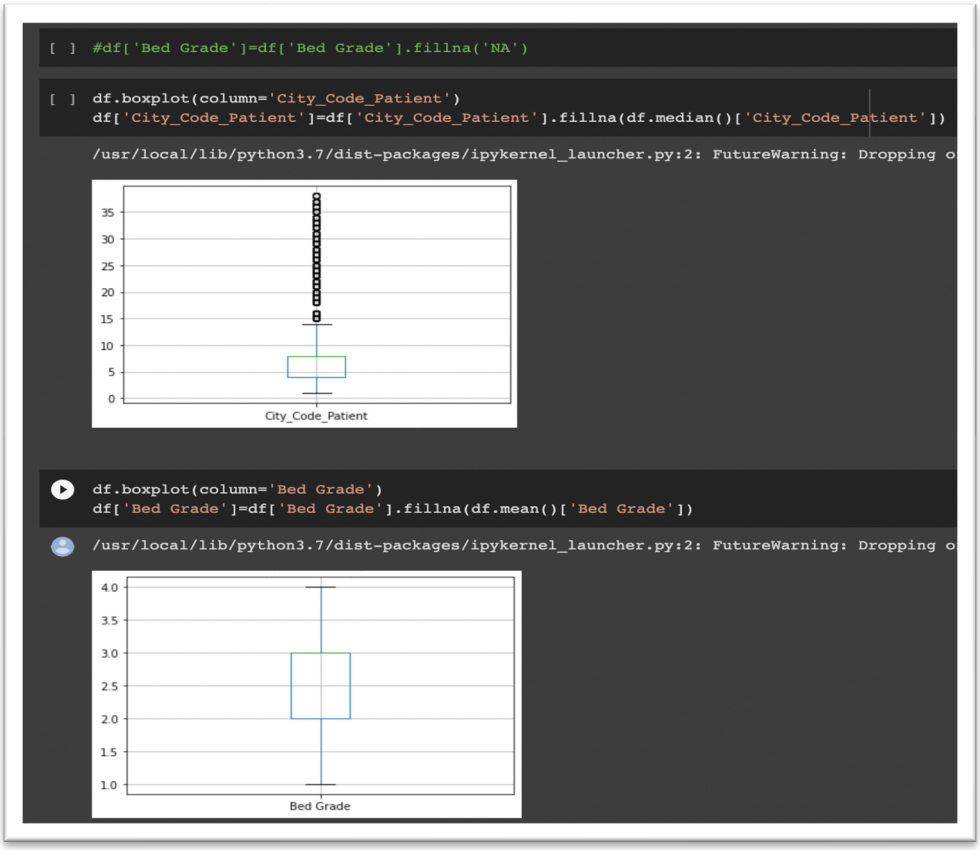
   I. **Missing Value Data Percentage** : In this dataset, only bed grade and city_code_patient give the missing value percentage of 0.0004 and 0.0142, respectively. Here I attached the code for your reference.[8]



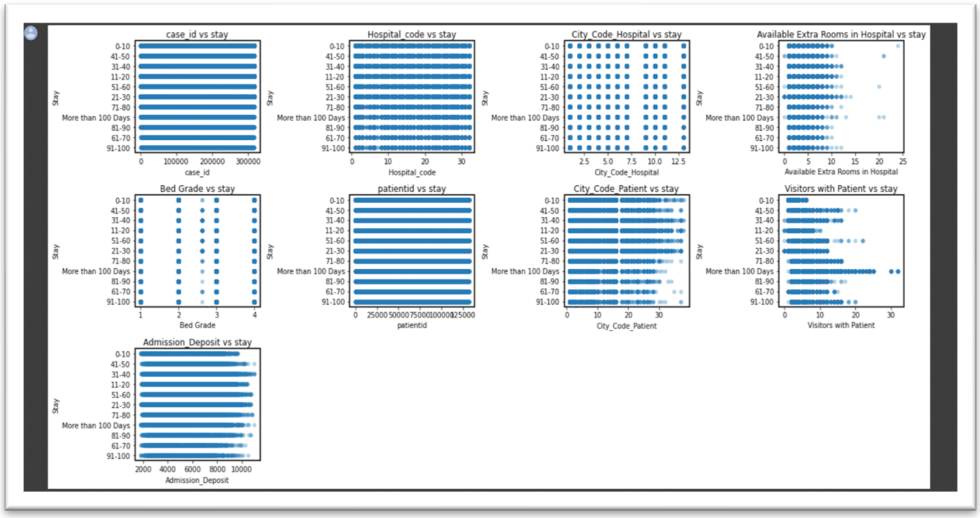
   II. **Handling Missing Value** : Now there are two ways available to handle the missing value,
      1) **Delete the Missing Value** : The benefit of this strategy is that it is a quick technique to handle the missing values, but this method works based on the missing values; if the missing value whole data represents either pattern or structure, then we can't simply delete the missing value, whereas if the missing values are the part of the random data, then if the row has the missing value, we can delete the entire row, same is applicable for the

column. Still, the disadvantage of this method is that sometimes we can delete valuable data from the dataset.[8]

2) **Imputing the Missing Value**: There are different ways available to fill in the missing values,
1) Replacing with arbitrary value: In this case, we can put any random value for missing values, but generally, we take values between the highest and lowest number available in the data attribute.
2) Replacing with mean value: This is the most common method used by the developer to fill in the missing values.
3) Replacing with mode: This method is used in the categorical value data set.
4) Replacing with the previous fill: for putting the missing value, we point the forward value and put it into the missing value place.[8]



III **Fetching the Numerical Data** : Here we write the code for finding the numerical Data, and visualize using the scatter_plot python library. [1]



IV **Processing the Categorical Data and Encoding** : Here the Hospital_type_code, Hospital_region_code,Department,Ward_Type,Ward_Facility_Code, Stay, Age, Type of Admission, Severity of Illness, so these are the categorical data so we Convert into categorical data format, after then we encoded it.

```
{'Hospital_type_code': array(['c', 'e', 'b', 'a', 'f', 'd', 'g'], dtype=object),
 'Hospital_region_code': array(['Z', 'X', 'Y'], dtype=object),
 'Department': array(['radiotherapy', 'anesthesia', 'gynecology', 'TB & Chest disease',
        'surgery'], dtype=object),
 'Ward_Type': array(['R', 'S', 'Q', 'P', 'T', 'U'], dtype=object),
 'Ward_Facility_Code': array(['F', 'E', 'D', 'B', 'A', 'C'], dtype=object),
 'Type of Admission': array(['Emergency', 'Trauma', 'Urgent'], dtype=object),
 'Severity of Illness': array(['Extreme', 'Moderate', 'Minor'], dtype=object),
 'Age': array(['51-60', '71-80', '31-40', '41-50', '81-90', '61-70', '21-30',
        '11-20', '0-10', '91-100'], dtype=object),
 'Stay': array(['0-10', '41-50', '31-40', '11-20', '51-60', '21-30', '71-80',
        'More than 100 Days', '81-90', '61-70', '91-100'], dtype=object)}
```

3. **Data Processing** : We are trying to find the desired output from the raw data using machine learning and artificial intelligence algorithms in this technique. Various types of data preprocessing methods are available based on the source of the data and what kind of output the user wants.[9]

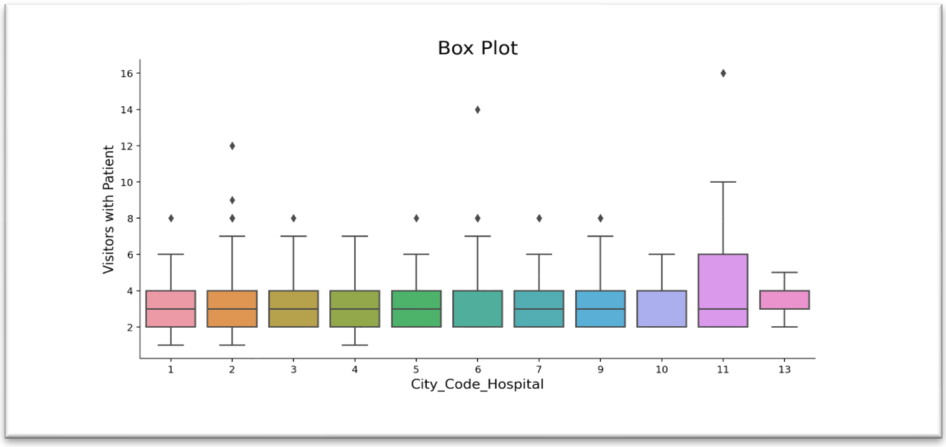
1) Batch Processing : In this type, all the data is collected; then it processed through batches; the example of this type is the payroll system because this type usually works on a large amount of the data.

2) Real-time Processing : Here the data is processed in real-time in seconds, so when the user gives the input data, then it starts processing and generates the desired output; because of that, it can't be possible for large data sets, so usually it is applicable for the only small amount of data. The best example for this method is the ATM machine, when users perform any kind of operation like debit or credit, it is directly reflected results into user's bank account.

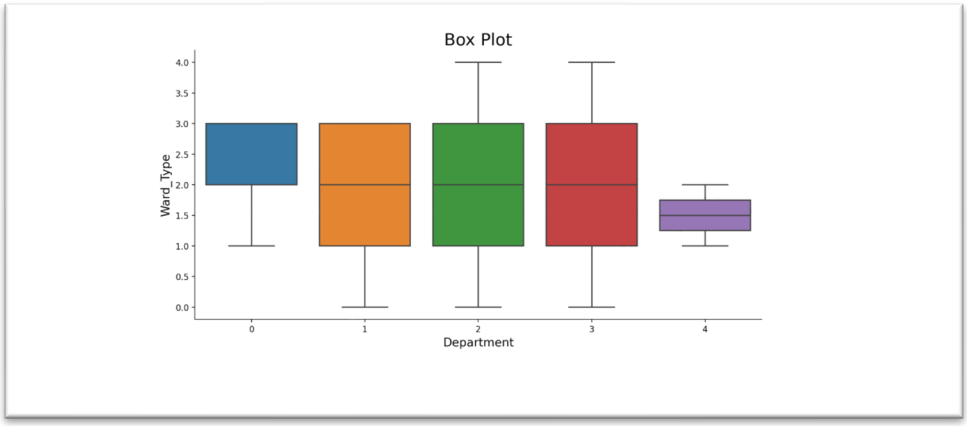
3) Online Processing : It is in some way like real-time processing; in this case, when the data is available, it is immediately transferred to the CPU for execution, the barcode scanning is a type of online processing example.

4. **Data Visualization** : Data visualization helps represent information in a graphically format, analyzing patterns and trends, and making business making decisions. In the below section I have attached the results, which I got from data, and at the end of report, I have included the conclusion section for overall summary of the data.

1) Box Plot : The box plot is helpful for representing numerical data. A box plot consists of 5 things in their visualization part, like Minimum, First Quartile (equal to 25 %), Median (Second Quartile), Third Quartile, and Maximum. This is useful for detecting the outlier data set.[11]
   - Here I have compared the City_Code_Hospital (X-axis) to the Visitors with Patient (Y- axis) , from we graph we conclude that the City_Code_Hospital with 11, have the highest number of visitor patient more than 10. In this visualization, the 16 is the outlier data.[3]

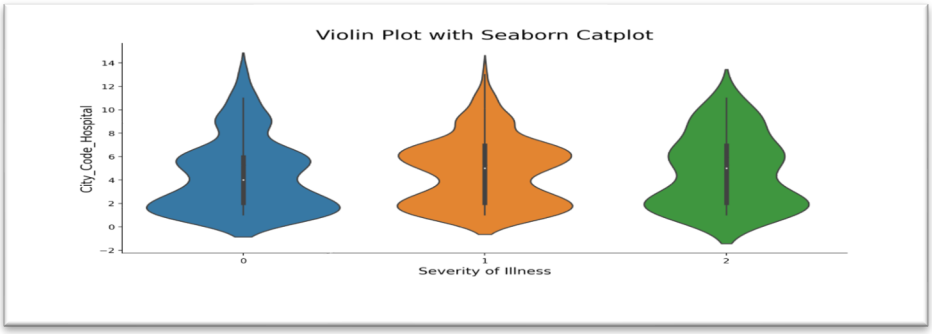


- Here I have attached the department (X-axis) to the Ward_type (Y-axis), we found that department zero don't have the second quartile and maximum values with respect to Ward_type.

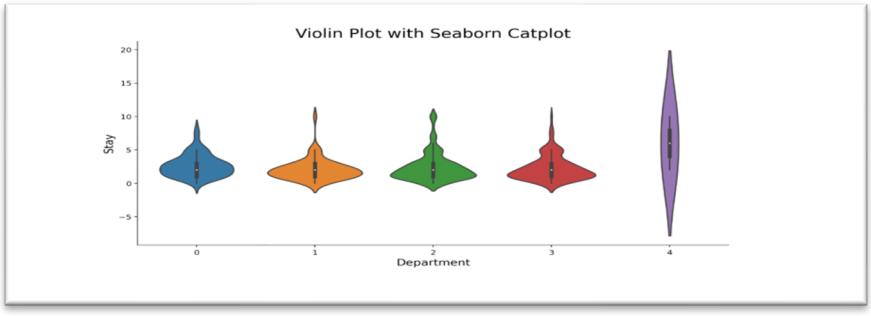


2) Violin Plot : The violin plot is some extent, like the box plot. Also, it helps show the quantitative data across one or more than one categorical variable. As it takes multiple data simultaneously, it is attractive and practical.[12]

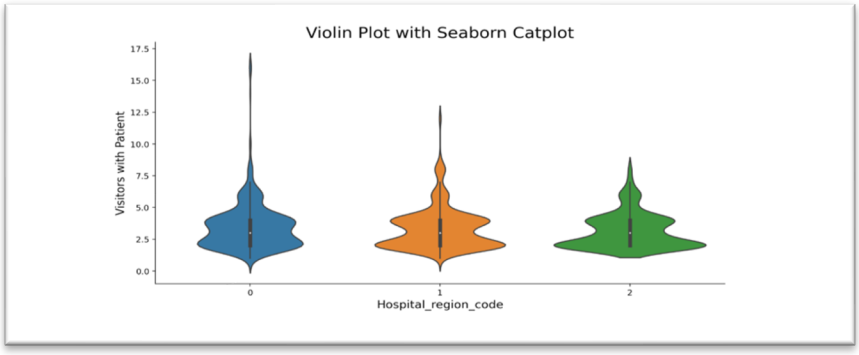
- Here I attached the City_Code_Hospital (Y – axis) with Severity of Illness (X- axis)



- Here I attached Department (X – axis) with Stay (Y- Axis)

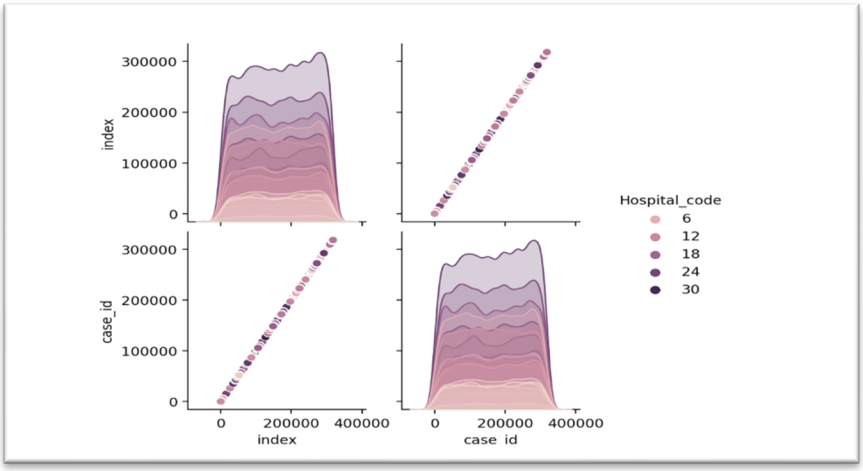


- Here I attached Hospital_region_code (X – axis) with Visitor with Patient (Y – axis)

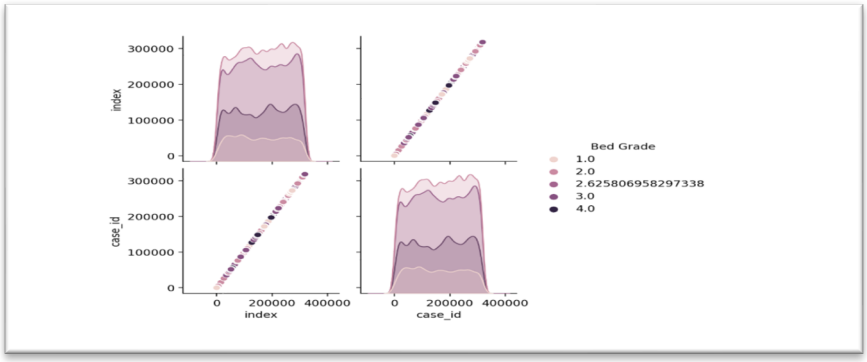


3) Pairplot : The pairplot visualization technique uses the pairwise relationship variables in the datasets; we can combine a large amount of data into a single figure.[5]

- We draw the pairplot between the case_id and Hospital_code

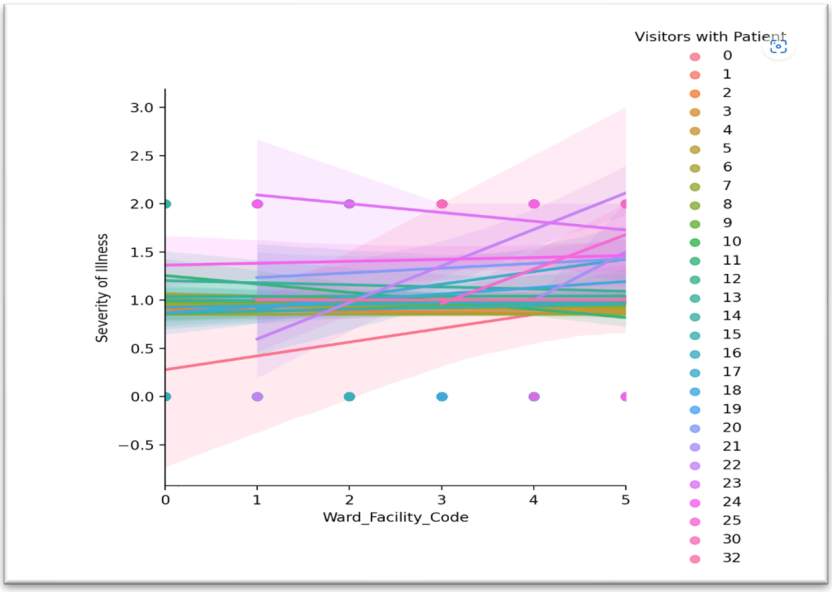


- Now the case_id with Bed Grade draw the pairplot with seaborn.

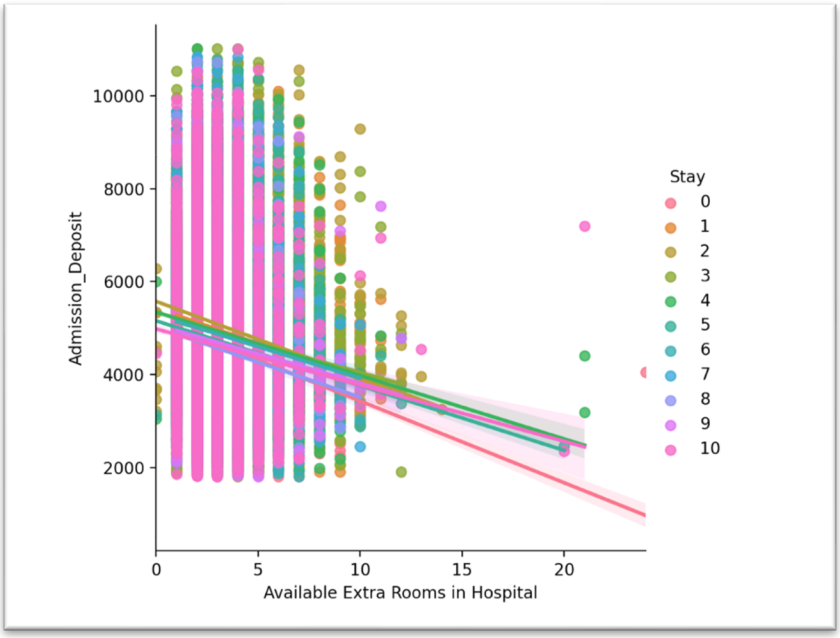


4) Lmplot : lmplot method is used for scatter plot visualization on the Facetgrid.[14]

- Ward_Facility_Code (X - axis) with Severity of illness (Y – axis) With taking as visitors with the patient

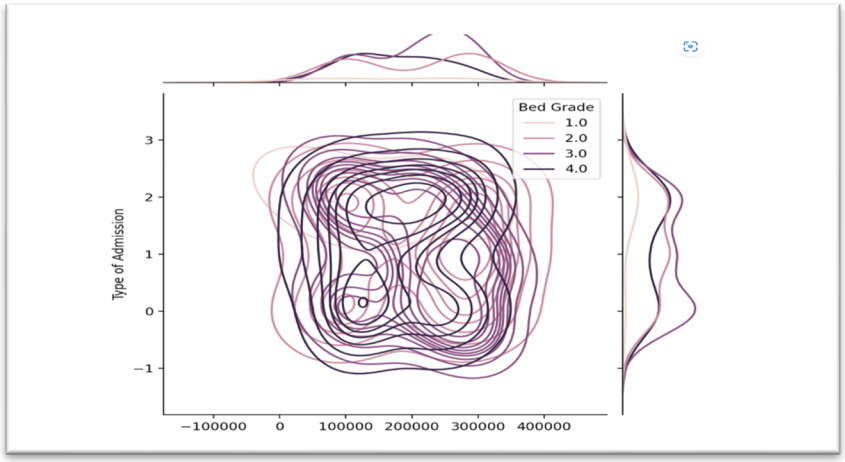


- Available Extra Rooms in Hospital (X – axis) with Admission Deposit (Y – axis) with Stay

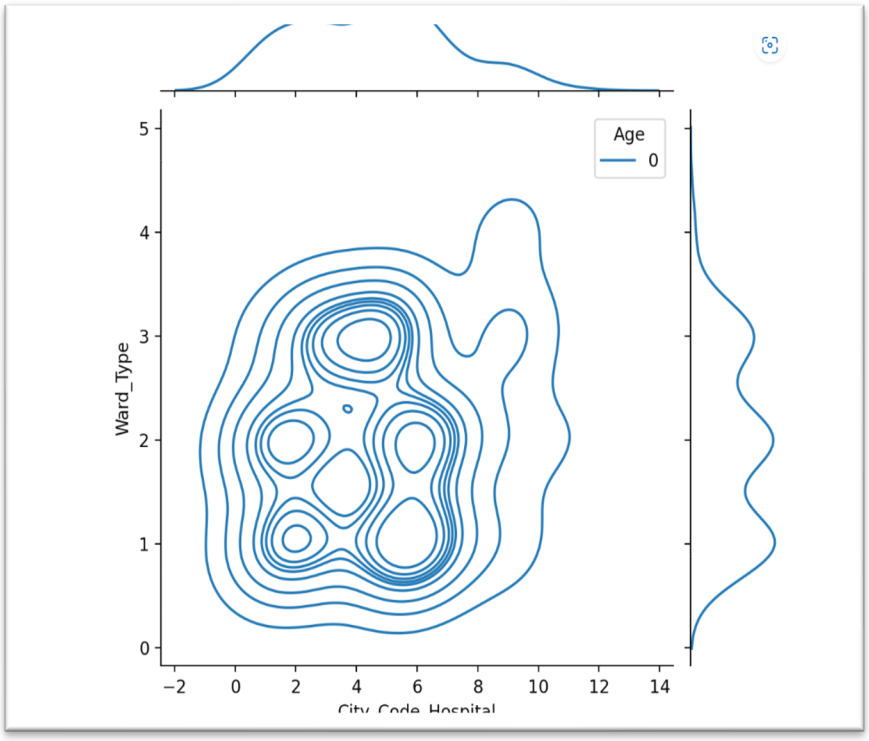


5) Displot : The Distplot is used to visualize a univariant observation set through the histogram.[15]

- The Case_id (X – Axis) and Type_of_Admission (Y- axis), now the Bed Grade represent the Hue for Box plot, Here I mentioned the visualize part for it.



- The City_Code_Hospital (X – Axis) and Ward_Type (Y- Axis), and the Age represent the Hue of Box plot.



Section III: Literature Review:

Apache Spark : Apache Spark is a framework for data processing that can swiftly conduct operations on substantial data sets and distribute functions over several computers, alone or in conjunction with other tools for distributed computing.[16]

- Since its 2009 inception in the AMPLab at U.C. Berkeley, Apache Spark has become distributed extensive data processing framework globally. The computer languages Java, Scala, Python, and R all have native bindings for Spark, which also supports SQL, streaming data, machine learning, and graph analysis.
- The Spark is written into the Scala and runs onto the java virtual machine (JVM).
- From the past comparison, it is found that Spark Is run programs up to 100 times faster than Hadoop map-reduce and ten times faster than disk. (class reference)
- In large datasets, speed is the main issue, but Apache spark has the rate to run a computation in memory.
- Spark covers a wide range of workloads in the same engine, making it easy to combine the different processing.

Apache Spark Components :

1) Mlib : Apache Spark includes libraries for machine learning and graph analysis methods on large amounts of data.[16]

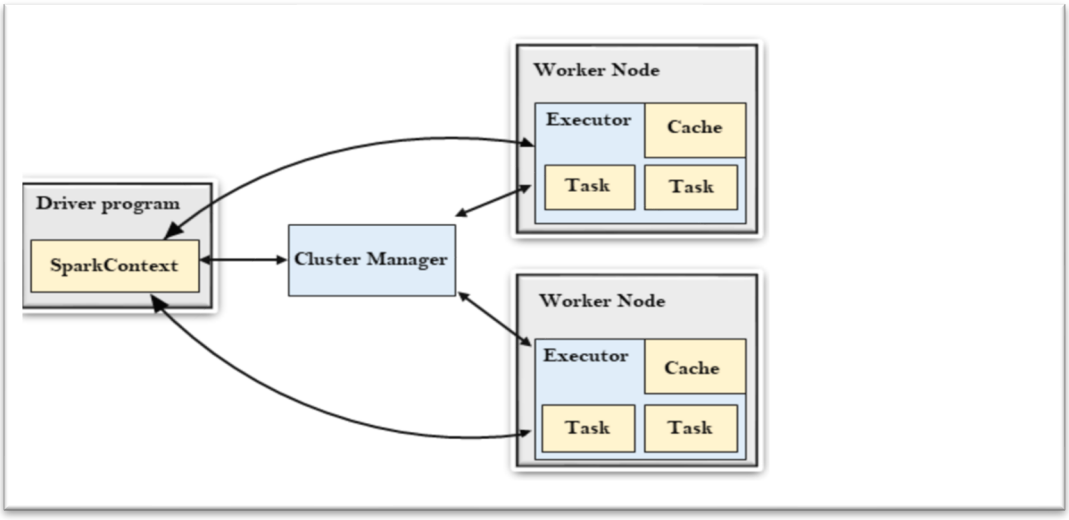
    - To create the machine learning pipelines, In Mlib offers easy feature extraction implementation for any structured data.
    - We can use the R or python language for training purposes, and later, it is saved using the Mlib; in the following steps, either java or Scala pipeline, we can use it.
    - Spark Mlib include all machine learning model technique like classification, regression, clustering, and so on.

Apache Spark Architecture : The master-slave architecture is used by Apache Spark.

- The first one is Resilient Distributed Dataset (RDD) and second is Directed Acyclic Graph (DAG)

1) Resilient Distributed Dataset : The Resilient Distributed datasets name is divided into three parts; the first one is Resilient, meaning that during the data computation if the data is a failure, it can automatically restore to it. And the meaning Distributed is distributing the data to different nodes, and the last term is datasets, which means the group of the data.[16]

2) Directed Acyclic Graph : DAG is useful for performing the sequence of computation on data. Here I attach the image for further reference and explained in detail below.



1) Driver Manager : The driver program creates the spark context object, and the main aim is to coordinate to the spark application, Running the spark context on a cluster perform the tasks ,First it executes nodes in cluster, then its application code the executors.
2) Cluster manager : The aim to is cluster manager to allocate the resource, but the spark is capable for running the larger data sets.
3) Worker Node : The role of worker node is to run application code into the cluster.
4) Exeutor : The executor launches the application on worker node It read and write data. Through the external sources. And it Is fact that every application have its own executor

Difference Between Spark Mlib vs H20 ?

- The main difference is that the H20 works on the .hex data format, while the Mlib takes data either RDD data structure format or Data structure of the spark,

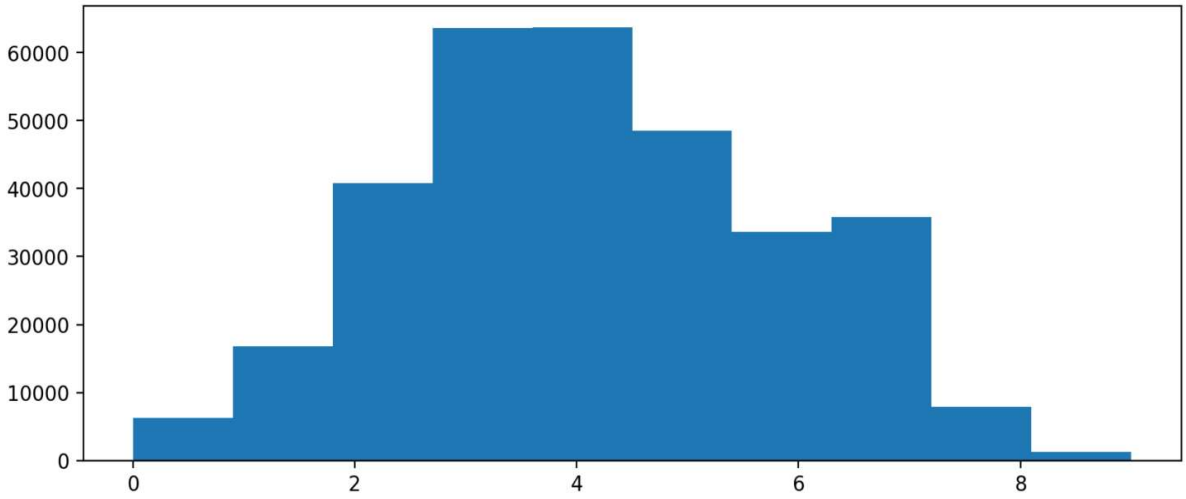
- From taking the different data into above format, the data management technique is required to solve that,

- Now from the Implementation purpose, we need a different implementation technique to implement for solving the problem.

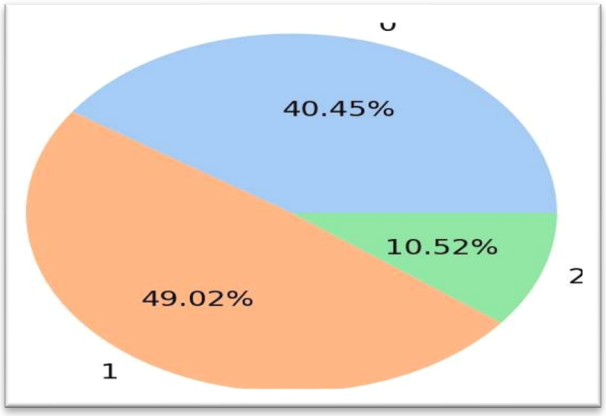
- In terms of the development workflow, spark ends up with the data bricks problem.

Section IV: Conclusion: Here I write the whole conclusion of the project along with the visualization part.

1) **Age Wise Analysis** :



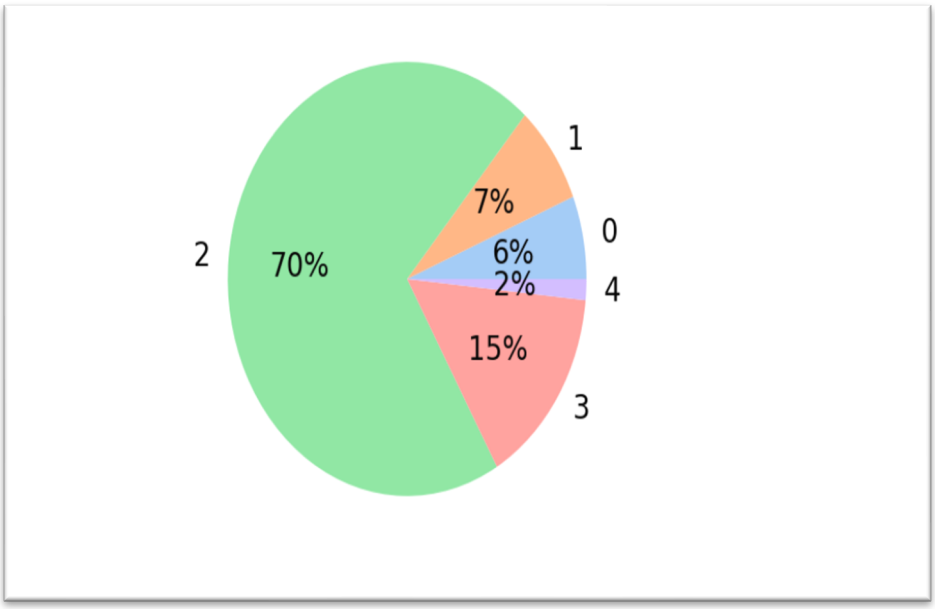
- From the above chart, we can conclude that the most of patients are found between the age 31-40 and 41-50 groups.
- So, in the next part of conclusion, I take these age groups.
- Now the problem we are facing is waiting in the hospitals, so to resolve this issue, we take the severity of illness with the department cases.



- From above we can say that each age groups have a moderate severity case compared to the dominant to extreme and minor cases.
- Especially, for the age group people between the 30-40 and 40-50 have moderate severity of illness, which is almost 3/5 th from the total cases.

- From that, the medical institution needs to focus on those cases, so from my point of view a person, with moderate level illness, can treated as soon as possible, as a result the overall length of the stay is reduced, and this will be helpful for other patient, who actually need to utilize the hospital services.

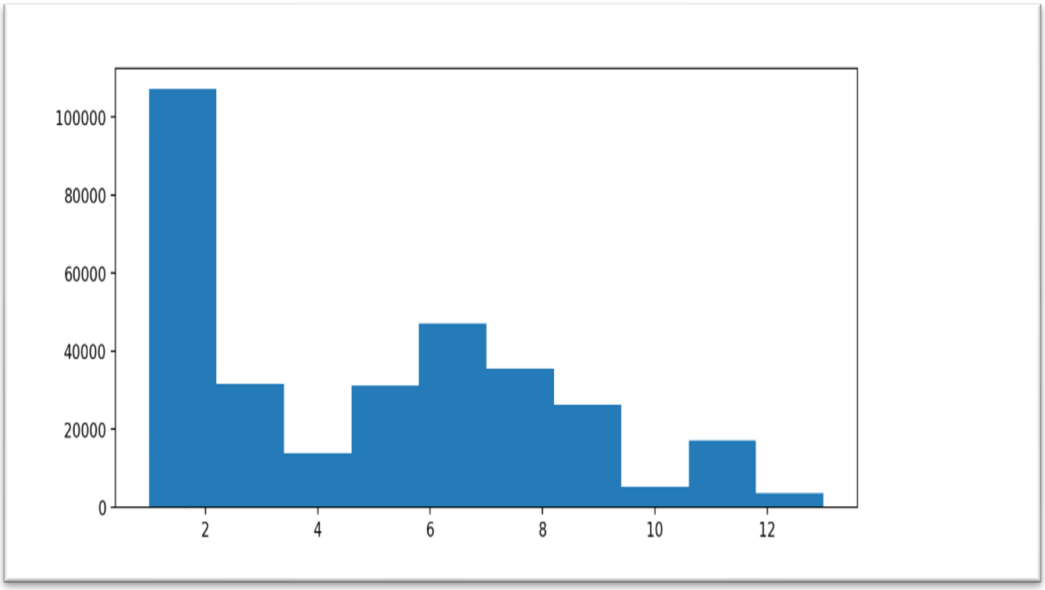
2) Percentage for each department for 0 age group



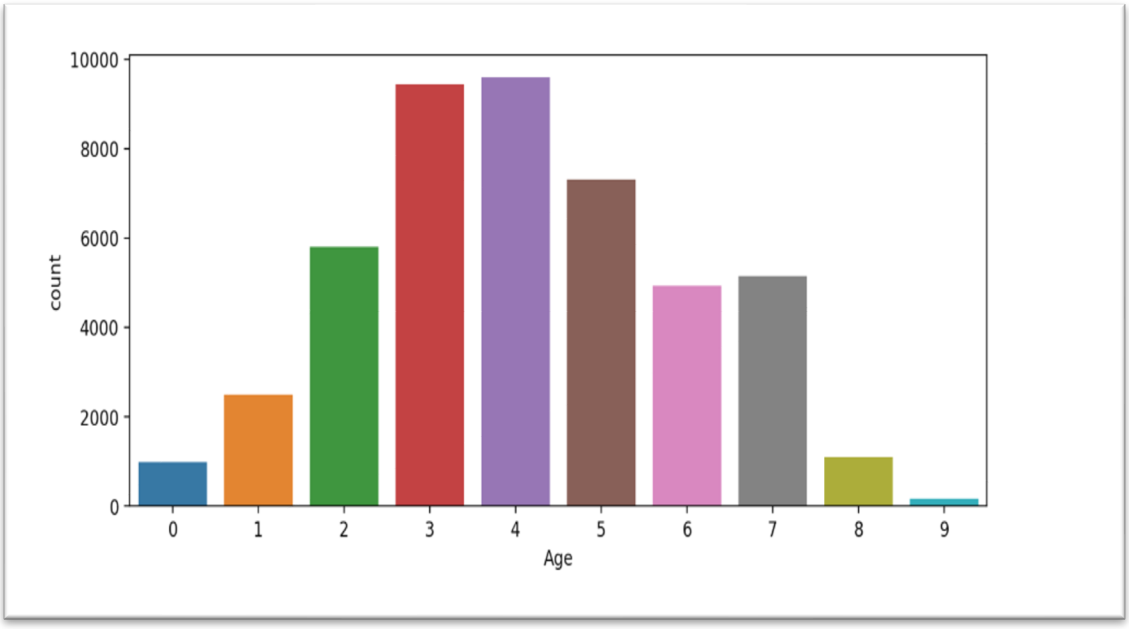
3) Department Analysis :

- I found that most of cases from all the ages groups are facing more gynecology, so the gynecology department improvement will reduce the overall cases.
- Also, there is second reason is that gynecology disease may be spread through the mass infection.

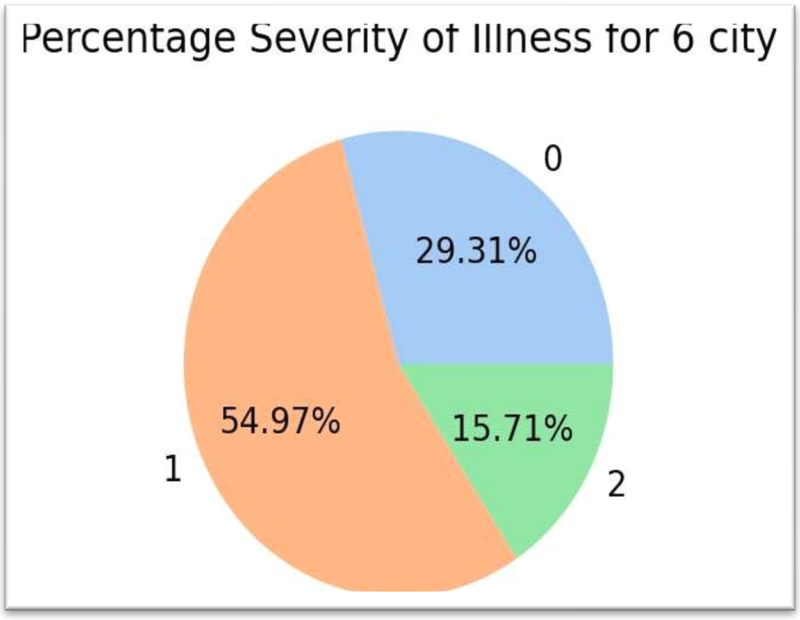
4) **City Wise Analysis** :



5) **Age wise distribution for '6' city**:



6) Percentage Severity of Illness for 6 cities:



- From above distribution chart, we can see most of the cased are belong to city '2'

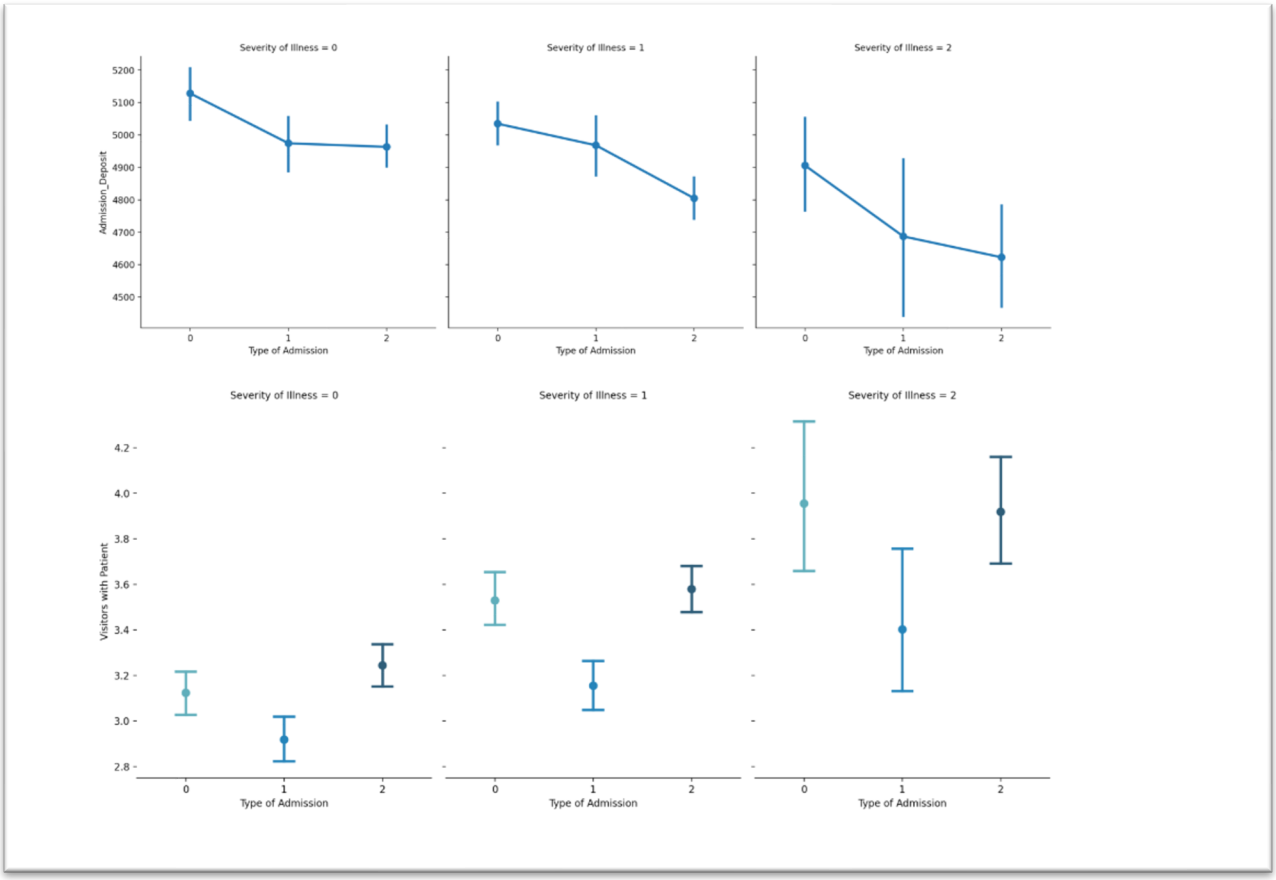
Reason :

- May not having good hospital facility
- Mass infection for city

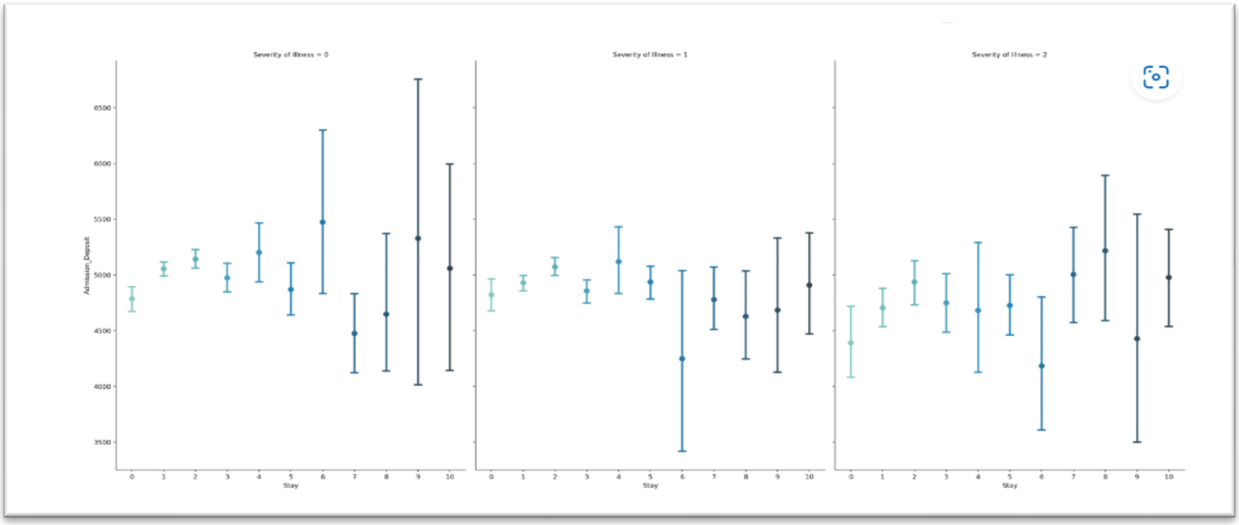
Solution :

- The situation can be tackled by improving the infrastructure for particular city.

7) Data Analysis For ‘0’ Age Range:



EDA With Respect To Stay Length :



- From the above graph, we can conclude that, irrespective of the proportion moderate severity cased, here the amount of stay for category is less , and the hospitalization cost is less for the severity of illness.
- If the person has some moderate of illness, want to stay for longer duration, the person needs to spend the more money like deposit, whereas the person, who has extreme moderated decease lower compared to other group.
- On the worst thing about, there are extreme case, and the amount of visitor for them is higher, so as a result it risks the overall infection rate, so to overcome from that hospital need to improve the facility.

Section V : REFERENCE:

1) https://docs.python.org/3/library/
2) https://www.kaggle.com/datasets/vetrirah/av-healthcare2?select=train.csv
3) https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.boxplot.html
4) https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.violinplot.html
5) https://seaborn.pydata.org/generated/seaborn.pairplot.html
6) https://seaborn.pydata.org/generated/seaborn.lmplot.html
7) https://www.javatpoint.com/machine-learning
8) https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/
9) https://www.simplilearn.com/what-is-data-processing-article
10) https://www.gartner.com/en/marketing/glossary/data-visualization
11) https://www.geeksforgeeks.org/box-plot-visualization-with-pandas-and-seaborn/
12) https://www.geeksforgeeks.org/violinplot-using-seaborn-in-python/
13) https://towardsdatascience.com/seaborn-pairplot-enhance-your-data-understanding-with-a-single-plot-bf2f44524b22
14) https://www.geeksforgeeks.org/python-seaborn-lmplot-method/
15) https://www.geeksforgeeks.org/seaborn-distribution-plots/
16) https://stackoverflow.com/questions/41054025/difference-between-h2o-ai-and-sparkmllib-from-machine-learning-algorithm-point-o
17) https://www.javatpoint.com/apache-spark-architecture

Group Member's Contribution:

- Akash: Mainly focused on the data cleaning and analyzing part of the project.
- Andrew: Handling missing values in the dataset.
- Henil: Worked on visualizing the dataset using various libraries.
  We combined our efforts and started working on the documentation of the final report.