Summer Internship Project Report on

Prediction and Optimization of Healthcare Resource Management for COVID 19 patients

In Partial Fulfillment of

PGDM-BDA *–* Batch-01

Submitted to

Prof. Sunita Daniel

Submitted by: Swapnil Malik Roll No. 015016



FORE School of Management, New Delhi

Title Fly

**SIP Progress Report**

Roll No.: 015016

Name of the student: Swapnil Malik

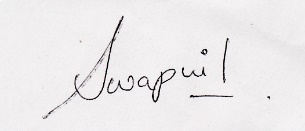
Name of the Faculty Guide: Sunita Daniel

**Meeting-1**

Date of Meeting: May 3, 2021

Topic/Work Discussed: Discussion of the domain in which the project is to be

pursued.

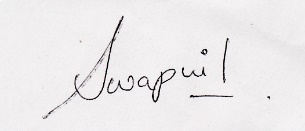


Signature of Student Signature of Faculty Guide

**Meeting-2**

Date of Meeting: May 14, 2021

Topic/Work Discussed: Discussion on the topic of the project.

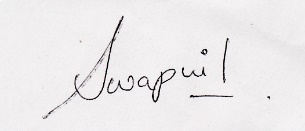


Signature of Student Signature of Faculty Guide

**Meeting-3**

Date of Meeting: May 13, 2021

Topic/Work Discussed: Meeting to finalise the topic.

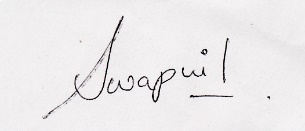


Signature of Student Signature of Faculty Guide

**Meeting-4**

Date of Meeting: June 2, 2021

Topic/Work Discussed: Discussion about collection of data.



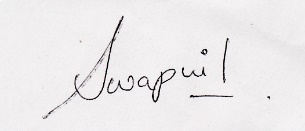
Signature of Student Signature of Faculty Guide

**Meeting-5**

Date of Meeting: June 5, 2021

Topic/Work Discussed: Updating about the progress. Discussed preliminary data

analysis.



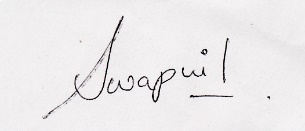
Signature of Student Signature of Faculty Guide

**Meeting-6**

Date of Meeting: June 12, 2021

Topic/Work Discussed: Discussion about the problems being faced in handling the

dataset.



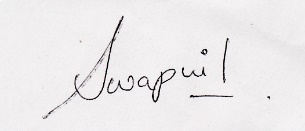
Signature of Student Signature of Faculty Guide

**Meeting-7**

Date of Meeting: June 19, 2021

Topic/Work Discussed: Updating about the progress. Discussion about the models

used in project code.

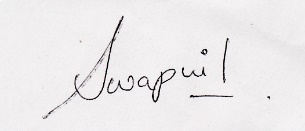


Signature of Student Signature of Faculty Guide

**Meeting-9**

Date of Meeting: June 26, 2021

Topic/Work Discussed: Updating about the progress and finalising the code.

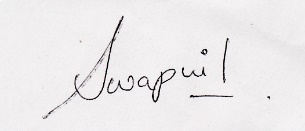


Signature of Student Signature of Faculty Guide

**Meeting-9**

Date of Meeting: July,1 2021

Topic/Work Discussed: Working in the accuracy of data



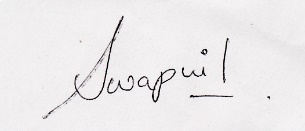
Signature of Student Signature of Faculty Guide

**Meeting-10**

Date of Meeting: July 12, 2021

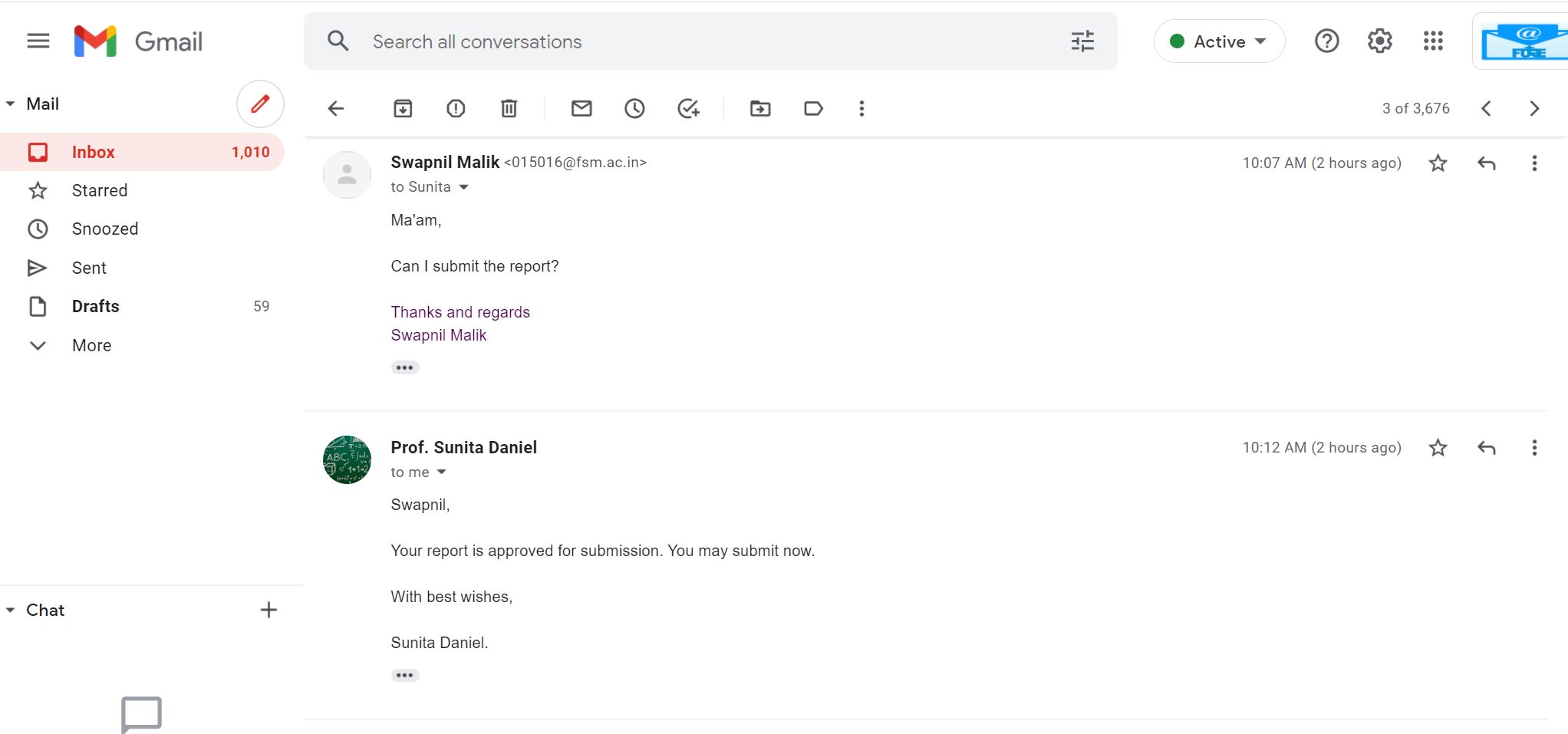
Topic/Work Discussed: Discussion on the first draft of the report. Changes suggest

by the faculty in charge.

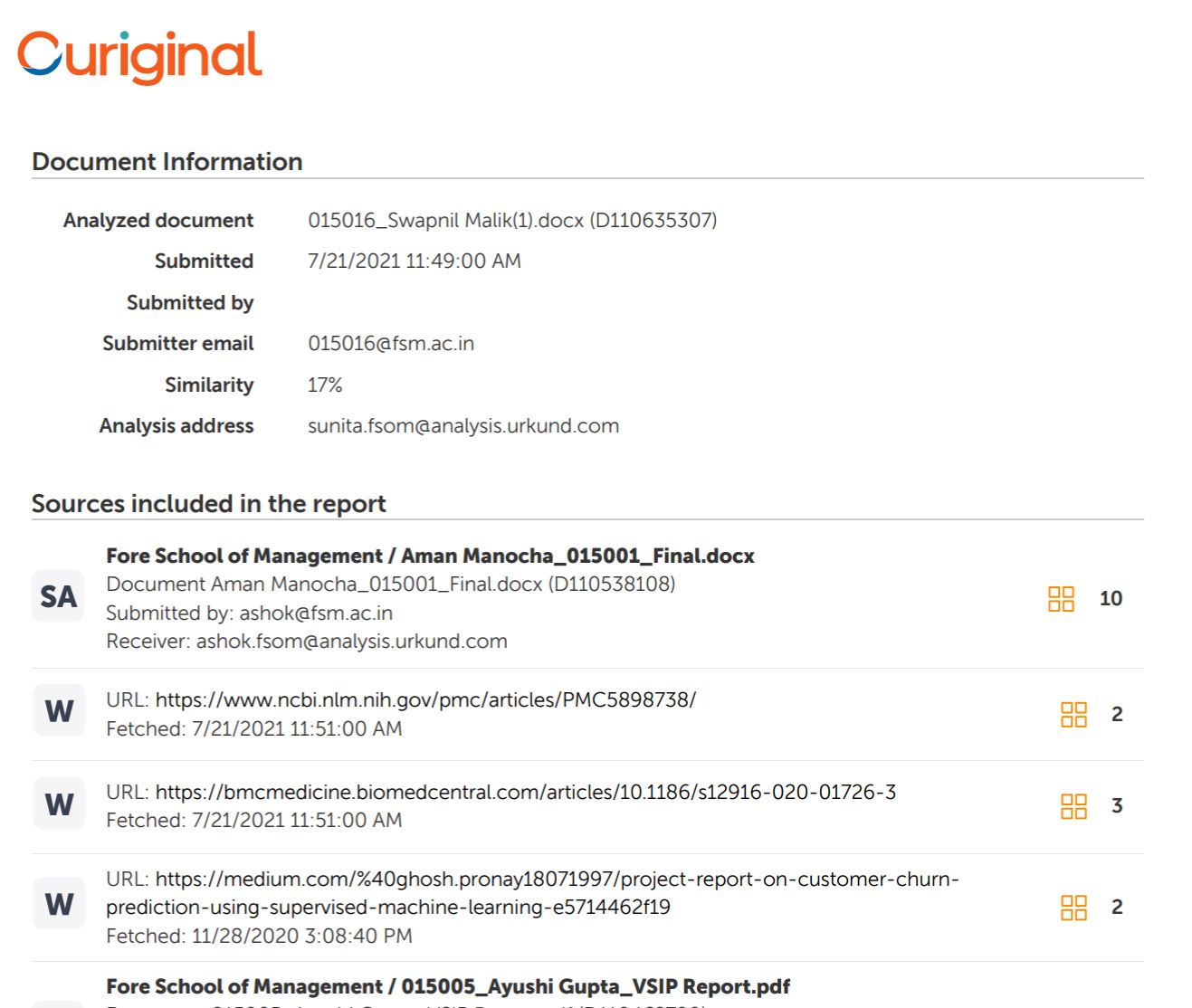


Signature of Student Signature of Faculty Guide

**Approval Mail**

****

**Urkund Analysis**

****

**Acknowledgement**

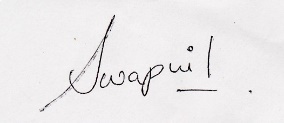
I would like to express my deep and sincere appreciation to my research guide, Dr. Sunita Daniel, Associate Professor, Information and Technology area, FORE School of Management, New Delhi for providing me the golden opportunity to carry out this research work and guidance throughout the research. Her vision, enthusiasm, sincerity and immense knowledge have deeply inspired me. Her constant stand to bring better and more insights has made this research more valuable. She focused on minute things encouraged me to prepare this report as clearly as possible.

I am enormously grateful for her help and support and without her supervision this research would not have been possible.

Declaration by the Student

I am, Ms. Swapnil Malik , Roll No. 015016 have completed my summer internship at FORE School Of Management and has submitted this project report entitled towards Prediction and Optimization of Healthcare Resource Management for COVID patient’s partial fulfilment of the requirements for the award of the Post Graduate Diploma in Management (FMG-29/IMG-14/FM-03/BDA-01) 2020-2022.

This Report is the result of my own work, no part of it has earlier comprised any other report, monograph, dissertation or book.



Signature

(Swapnil Malik)

Certificate

This is to certify that Ms. Swapnil Malik, Roll No. 015016 has completed his/her summer internship at FORE School Of Management, Delhi and has submitted this project report entitled Prediction and Optimization of Healthcare Resource Management for COVID patients towards partial fulfilment of the requirements for the award of the Post Graduate Diploma in Management BDA-01 2020-2022.

This Report is the result of her own work and to the best of my knowledge, no part of it has earlier comprised any other report, monograph, dissertation or book. This project was carried out under my overall supervision.

Date: 16th July, 2021

————————————-

Internal Faculty Guide

Executive Summary

This study deals with Prediction of Length of Stay at hospitals at hospitals. It helps to optimize resources management and have efficient allocation of it. Because of the COVID 19 hospitals have faced a major issue of lack of bed facility. While the available resources are limited an attempt could be made to make the most out of scarce resource. Forecasting length of stay could help hospital to better make informed decision leading to smoother operations.

For the study patient data of various hospitals was taken from Kaggle. It consisted of 318438 rows and 18 columns which needed to pre-processed and cleansed according to study. Null values were removed from 2 columns i.e. Bed Grade and patent id by taking the highest of the 2. Using Matplotlib, Plotly and Seaborn libraries the given data was visualised and insights were drawnabout admission deposit, age, hospital region, illness severity, number of visitors had an impact on stay days at hospital.

Since most of data was categorical methods like Chi square testing was done to determine relationship between categorical variable and stay days at hospital. Label encoder encoded categorical columns to numerical.

For prediction three models are used KNN Classifier, Random Forest Classifier and Decision Tree Classifier. The best fit model was found to be Random Forest Classifier.

While the study tried to eliminate the any limitation but still failed to in capture external factors, patient’s financial condition and expert opinion.

This study aims to be used in by the hospitality business for their future proposals and also, help deal with the COVID time.

Table of Contents

* **Acknowledgement**
* **Declaration**
* **Certificate**
* **Table of Contents**
* **Executive Summary**

1. **Chapter 1: Introduction**
   1. Background of the Study
   2. Literature Review
   3. Purpose of the Study
   4. Relevance of the Study
   5. Objectives of the Study
2. **Chapter 2: Methodology**
   1. Project Design
   2. Universe of the Study
   3. Locale of the Study
   4. Data Collection Method
3. **Chapter:3 Data Analysis**
   1. Data Exploration
   2. Data Pre-Processing
   3. Data Visualisation

3.3.1. General Visualization

3.3.2. Stay Days Analysis

3.3.3. Categorical Data Analysis

* 1. Data Engineering
     1. Label Encoding
     2. Stay Days: splitting into classes
  2. Model Building
     1. Splitting the Dataset
     2. Modelling

1. **Chapter 4: Result and Discussions**
2. **Chapter 5: Limitation and Scope**
   1. Limitation
   2. Scope
3. **Biblography**
4. **References**

**8 Annexure A.**

Application used

Libraries Imported

**Table of figures**

* Figure 1: Head of the data
* Figure 2: Info
* Figure 3: Null Value
* Figure 4: Bed\_Grade maximum value code
* Figure 5: Bed\_Grade missing value filled
* Figure 6: City\_Code\_Patient maximum value code
* Figure 7: City\_Code\_Patient missing value filled
* Figure 8: Bar Graph showing hospital type with patient density in hospital
* Figure 9: Bar Graph showing Age and count of people admitted
* Figure 10: Boxplot showing Available\_Extra\_Rooms\_in\_Hospitals, Patient\_Visitors and Admission\_Deposit
* Figure 11: Categorization of hospital type by region
* Figure 12: Correlation plot of all the attributes of data
* Figure 13: Bar graph here represents count of patients admitted in different categories of stay days
* Figure 14: Count of category of stay and type of admission
* Figure 15: Bar chart showing category of stay and type of admission
* Figure 16: Stay Days and amount Admission\_deposit are depicted by Boxplot
* Figure 17: Stay Days and amount Admission\_deposit are depicted by Bar graph
* Figure 19: Subplot analysing the categorical variables
* Figure20: Two-way Chi Square test for the relationship between Stay\_Days and Hospital\_type
* Figure 21: Chi square dependency results
* Figure 22: Two-way Chi Square test for the relationship between Stay\_Days and Hospital
* Figure 23: Chi square dependency results
* Figure 24: Two-way Chi Square test for the relationship between Stay\_Days and Hospital
* Figure 25: Chi square dependency results
* Figure 26 : Two-way Chi Square test for the relationship between Stay\_Days and Hospital\_Region
* Figure 27 : Chi square dependency results
* Figure 28 : Two-way Chi Square test for the relationship between Stay\_Days and Department
* Figure 29 : Chi square dependency results
* Figure 30 : Two-way Chi Square test for the relationship between Stay\_Days and Ward\_Type
* Figure 31 : Chi square dependency results
* Figure 32 : Chi square dependency results
* Figure 33 : Two-way Chi Square test for the relationship between Stay\_Days and Ward\_Facility
* Figure 34 : Chi square dependency results
* Figure 35 : Two-way Chi Square test for the relationship between Stay\_Days and Bed\_Grade
* Figure 36 : Chi square dependency results
* Figure 37 : Two-way Chi Square test for the relationship between Stay\_Days and City\_Code\_Patient
* Figure 38 : Chi square dependency results
* Figure 39 : Two-way Chi Square test for the relationship between Stay\_Days and Type of Admission
* Figure 40 : Chi square dependency results
* Figure 41 : Two-way Chi Square test for the relationship between Stay\_Days and Illness Severity
* Figure 42 : Chi square dependency results
* Figure 43 : Two-way Chi Square test for the relationship between Stay\_Days and Age
* Figure 45: Label Encoding
* Figure 46 : Splitting into classes
* Figure 47 : Separated to X and Y where stay days was the target variable
* Figure 48 : 13 attributes of X
* Figure 49 : KNN Classifier accuracy
* Figure 50 : Decision Tree Classifier
* Figure 51 : Random Forest accuracy
* Figure 52 : Accuracy of the used models

**Chapter 1**

# **Introduction**

With the outbreak of pandemic the health infrastructure has crumbled worldwide. Newspapers, websites and WhatsApp groups are filled with demand for oxygen cylinders, vaccines, hospital beds, ICU etc.

India's second wave of coronavirus disease has been aggravated by significant stress on our already limited healthcare resources. Health experts had for long warned about the consequences of underfunding the country’s health infrastructure. COVID 19. But the recent surge has exposed India’s health preparedness even more than the most pessimistic forecasts. More than 2,000,000 people were reported positive for COVID, which is very high compared to health resources nationwide. Hospitals with intensive care facilities face the challenge of limited equipment availability.

To keep up with the required demand of healthcare resources the two way out are either to increase resources or optimally use the available resources. While the government works over the initial one hospitals can work over the later one.

Being able to analyse and introspect number of days required to could help optimally and efficiently utilize the resources.

The study focused on patient length of stay, which is an important indicator of the effectiveness of hospital management. Reducing hospital stays leads to a reduced risk of infection and medication side effects, improved quality of care, and increased hospital profitability with more efficient bed management.

1.1 Background of the study

India's medical infrastructure has faced unprecedented pressure, especially during the COVID 19 wave, with rapidly increasing demand for healthcare in hospitals, intensive care units (ICUs), hospital beds etc. With over 3.2 hospitals per 1000 people, India’s health infrastructure showed poor and alarming rates. As the pandemic progresses, the identification of the associated human health and the resources needed (beds, staff and equipment), has become a top priority for many countries. In the future, demand forecasting requires to give an estimate of how long COVID-19 patients in need of different levels of the health care system.

The length of hospital stay is an important parameter to monitor, as well as information to improve the effectiveness and efficiency of healthcare management. The prediction will make it easier for hospitals to identify patients who are at high risk of long-term hospital stays. For the identification of patients at high risk for long-term stay on their treatment plan can be optimized in order to minimize the length of stay and the likelihood of infection is low and the room was of visitors. In addition, a basic knowledge of the length of stay can be of help, supply chain management, so as to plan the distribution of a room and a bed. The challenge is to be professional and to optimize the management of the hospital's operations.

* 1. **Literature Review**

Feibert, Diana Cordes, Jacobsen, Peter, Wallin &Michael in their research stated that “Healthcare costs are rising due to an aging population and more sophisticated technologies and treatments. At the same time, patients expect high-quality care at an affordable cost. As a result, the healthcare industry is under increasing pressure to reduce the cost of healthcare delivery while delivering high-quality healthcare. Hospital logistics operations provide an important opportunity to control healthcare costs through the implementation of best practices.Based on the analysis and comparison of case studies, a set of factors influencing the decision to improve the logistics process of healthcare has been identified.”

Stefania Tattoni, Diana Giannico, Massimiliano Schiraldi & Massimiliano Schiraldi in their research stated that “The article focuses on identifying opportunities for improvement by analyzing processes within cleaning departments, and the case of the radiology department of a public hospital in Rome has been reported. Initially, a complete study of the internal processes of the analyzed service was carried out to identify the bottlenecks and the causes of the loss of service level. Second, to make better use of resources and increase service productivity, a method of optimizing resource management through patient planning and shift management has been identified: patients are divided into inpatients (admitted) and outpatient (which were admitted to the hospital only in tests and not requiring an overnight stay) and a higher use of resources was achieved through flexible management of the training process for trainers.

Nilmini Wickramasinghe, Rajeev K. Bali, M. Chris Gibbons, J. H.James Choi, Jonathan L in their research stated that “A systematic approach and application of knowledge management (KM) principles and tools can provide the basis for improving decision-making in healthcare. Boyd's combination of the OODA (Observation, Direct, Decide, Act) Loop and Continuous Intelligence provides an integrated, systematic, and dynamic model for ensuring that decision makers care Healthcare professionals always have the relevant and necessary knowledge that will help ensure that the outcomes of healthcare decision-making are optimized for maximum patient benefit. Examples of procedures in the orthopedic operating room will illustrate the application of the integrated model to support effective decision-making in a clinical setting.”

Pei-Fang (Jennifer) Tsai, Po-Chia Chen, Yen-You Chen, Hao-Yuan Song, Hsiu-Mei Lin, Fu-Man Lin, and Qiou-Pieng Huang in their research stated that “For admission management, the ability to predict length of stay (SD) from the pre-hospital stage can be useful for monitoring the quality of inpatient care. This study aims to develop artificial neural network (ANN) models to predict the life expectancy of hospitalized patients with one of three main diagnoses: coronary atherosclerosis (CAS), heart failure (HF) and acute myocardial infarction (AMI) in a cardiovascular unit at a Christian Hospital in Taipei, Taiwan.”

Ronald Lagoe, Barbara Drapola, Mary Luziani & Louise Pernisi in their research stated that “This study evaluated the reduction in length of stay for adult medication and surgery at combined hospitals in Syracuse, New York between 1998 and 2016. Study Based on the Severity System of illness for all patients. Using this approach, he controlled for changes in disease levels in the hospital population. Data from the study indicate that the reduction in adult medical and surgical stays in Syracuse hospitals between 1998 and 2012 resulted in a reduction in the number of additional annual days compared with the national average. severity-adjusted is 49,000, or the census daily mean is 134.2. The move to Medicare-initiated discharge reimbursement is believed to be the main driver of this cut.”

Ronald Lagoe, Louise Pernisi, Mary Luziani & Shelly Littau in their research stated that “The length of stay of unusual patients, who are hospitalized for long periods of time, presents a significant challenge to improving healthcare efficiency. This study identified unusual patients and their treatment programs in the Syracuse, New York metropolitan area. He showed that in 2013, out-of-province patients accounted for 2.4% of adult drug discharge and the average daily number exceeded 53.3 patients at Syracuse hospitals. Median length of stay of inpatients with post-hospital complications was nearly three times higher than in the inpatient/surgery population in 2013”

Aya Awad ,Mohamed Bader-El-Den & James McNicholas in their research stated that “The study focused on predicting measurable outcomes, including risk of complications, death, and length of hospital stay. Length of stay (SD) is an important metric for both healthcare providers and patients, influenced by many factors. In particular, length of stay in the intensive care unit is of great importance, both for patient experience and for the cost of care, and is influenced by factors that characterize the highly complex ICU environment. In addition, the paper presents a classification and evaluation of methods of length of stay analysis and mortality prediction related to a group of related research articles published from 1984 to 2016 related to to the field of survival analysis. In addition, the document highlights a number of gaps and challenges in this area.”

Kieran Stone, Reyer Zwiggelaar, Phil JonesNeil &Mac Parthaláin in their research stated that “The main goal of hospital managers is to plan and organize health care properly by arranging the facilities, equipment and human resources necessary for the operation of the hospital in accordance with the needs of the hospital. patient needs while minimizing health care costs. A new approach to service loss prediction is studied. In this article only data based on general patient diagnoses are used. This data is collected during the patient's hospital stay along with other general personal information such as age, gender, etc. Several different classifiers are used to understand the possibility of performing knowledge discovery on this limited dataset. They demonstrate about 75% classification accuracy. In addition, a range of other perspectives are explored to provide insight into the contribution of individual characteristics and how the findings can be used to influence decision-making, planning and planning, and the management of staff and organizations. resources.”

Christopher M. McDermottGregory Neal StockGregory Neal Stock in their research stated that “Purpose - As hospital costs continue to rise, increasing attention is being paid to how these institutions are and should be run. This attention often focuses on service costs, quality (often measured by mortality), and length of stay. Hospital leadership has many options to meet these challenges. To analyze these relationships, the paper uses data from the New York State hospital population. The performance measure was the MSD for the patient, adjusted for the composition and severity of cases in each hospital. However, the paper found a significant interaction effect between capital expenditure and salary and staff management perspective.”

# Evelene M Carter & Henry WW Potts in their research stated that “Investigated whether factors could be identified that significantly influence hospital stay from those available in an electronic patient record system, using total knee replacement as an example. To investigate whether a model could be generated to predict length of stay based on these factors to aid in resource planning and patient expectations regarding length of stay. Factors that had a significant effect on length of stay were age, gender, and consultant, destination of discharge, deprivation, and ethnicity.”

# Tayebeh Baniasadi, Kobra Kahnouji, Nasrin Davaridolatabadi & Saeed Hosseini Teshnizi in their research stated that “One of the performance indicators to determine the efficiency and optimal use of hospital resources is length of stay (SD). This study aimed to determine the length of hospital stay of patients and factors affecting the length of hospital stay of children. The results show that at this hospital, length of stay has a significant relationship with the variables of time of admission, place of residence, type of hospital admission and level of treating physician.”

Tayebeh Baniasadi, Marjan Ghazisaeedi, Mehdi Hassaniazad, Sharareh R. Niakan Kalhori, Mehraban Shah in their research stated that “Understanding each of the factors that influence length of stay, especially in surgical wards, can be important in planning the optimal use of hospital resources. This study aims to identify the factors affecting the length of stay (DS) in the surgical department and then to propose technological solutions.”

Takeru Abe in his research stated that “Reducing hospital stays leads to reduced risk of infections and drug side effects, improved quality of care, and increased hospital profitability with more efficient bed management. The aim of this study was to determine what factors are associated with length of hospital stay, based on electronic health records, to more effectively manage hospital stay.”

**1.3 Purpose of the study**

It has been noted that health infrastructure is facing extremely challenging situation in a densely populated country, India. The purpose of this study was to determine which factors are associated with length of hospital stay, in order to manage hospital stay more efficiently. This will provide an aid in logistic requirement based on historical data of numerous patients. The hospitals can predict length of stay with the help of data science based on relevant factors identified in the study. Lower Length of stay can help reduce chances of hospital acquired infections.

**1.4 Relevance of the study**

Impact on healthcare resources can help create strategies and provide real-time solutions for the requirement of hospital resources, such as personal protective equipment and ventilators. Data-driven government hospitals can identify new ways to respond to pandemics. With unprecedented time, the optimal use of limited resources can save lives. People can access on-demand resources given that they have been used effectively by hospitals in the past.

**1.5 Objectives of the Study**

* Predicting demand for hospital services and an estimate of how long each person will require hospital care.
* Logistics support such as planning the bed allocation and helping the hospital make optimal use of the equipment required by the patient.
* Identify factors which have an impact of Length of stay of patients.
* Analyze which model is best for prediction of Length of Stay.

**Chapter 2**

# **Methodology**

**2.1 Project Design**

For the purpose of study an exploratory research has been done to explore the factors that will affect the length of stay at hospitals. Jupiter notebook was used to analyse the data collected from Kaggle. The raw data constituted of 318438 rows and 18 columns which needed to pre-processed and cleansed according to study. Null values were removed from 2 columns i.e. Bed Grade and patent id by taking the highest of the 2 . Using Matplotlib, Plotly and Seaborn libraries the given data was visualised and insights were drawn thereon. The factors which affect a patients at hospitals were explored.

Chi square testing was done to determine relationship between 2 categorical variables. Since most of the data constituted of categorical variables it required label encoder to be able to fit to the model.

Data Engineering included Label Encoder which would encode categorical variables to numerical variables.

Also, dividing the data into various classes to get desired results.

The basic design of the project includes:

Three Machine Learning models were used i.e. Decision tree Classifier, Random Forest Classifier and K Nearest Neighbours Classifier.

* Decision Tree Classifier- A decision tree is a flowchart type tree structure where the inner node represents an attribute, the branch represents a decision rule, and each leaf node represents the result. A decision tree consists of
  + Nodes: Tests the value of a certain feature
  + Edges/Branch: Represents a decision rule and connects to the next node.
  + Leaf nodes: These are the terminal nodes represent class labels or a class distribution.
* Random Forest Classifier- Random forest is an aggregation method that combines many decision trees for classification. Hence, the results of the random forest are generally better than the results of the decision trees. The random forest is a supervised learning algorithm. It can be used for both classification and regression. It is also the most flexible and easy-to-use algorithm. As we know, a forest is made up of trees. The more trees there are, the stronger the forest. Random forests generate decision trees on randomly selected data samples, take predictions from each tree and then choose the best solution through voting. It also acts as a great indicator of feature importance.
* K Nearest Neighbour Classifier- KNearest Neighbor or KNN is a supervised nonlinear classification algorithm. KNN is a nonparametric algorithm, that is, it does not make any assumptions about the underlying data or its distribution. It came out to be one of the simplest and most widely used algorithms, it depends on its neighbours (K value) and is applied in many industries like the financial industry, healthcare industry, etc.

The models were compared on the basis of accuracy shown.

Further it includes determining scope and limitation of the project

## **2.2 Universe of the Study**

As per Institute Montaigne India has a total of 43,486 private hospitals, 1.18 million beds, 59,264 ICUs, and 29,631 ventilators. On the other hand, there are 25,778 public hospitals, 713,986 beds, 35,700 ICUs, and 17,850 ventilators. India’s population is 1.3 billion. Patients admitted to the hospitals are considered to be universe of study for this study.

## **2.3 Locale of the Study**

A total of 318,438 patient’s records of 32 anonymous hospitals were taken from into consideration.

**2.4 Data Collection Method**

The data was collected from secondary source, Kaggle. Source of data: https://www.kaggle.com/arashnic/covid19-hospital-treatment

**Chapter 3**

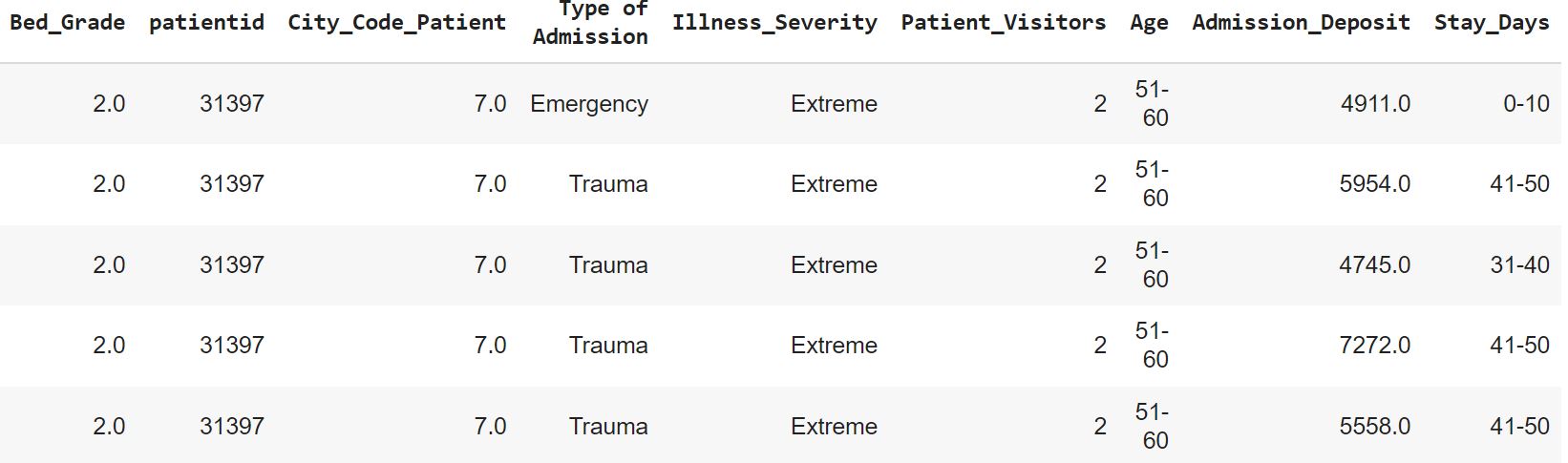
# **Data Analysis**

**3.1. Data Exploration**

This dataconsists of 318438 rows. It consisted of 18 attributes namely: ' Case id ', ' Hospital ', ' Hospital type', ' Hospital city', ' Hospital region', 'Available Extra rooms in Hospital', ' Department ', ' Ward Type ', ' Ward Facility', ' Bed Grade ', 'Patient id', 'City Code Patient ', 'Type of Admission', 'Illness Severity ', 'Patient Visitors', Age', 'Admission Deposit' and 'Stay Days'.

* Head

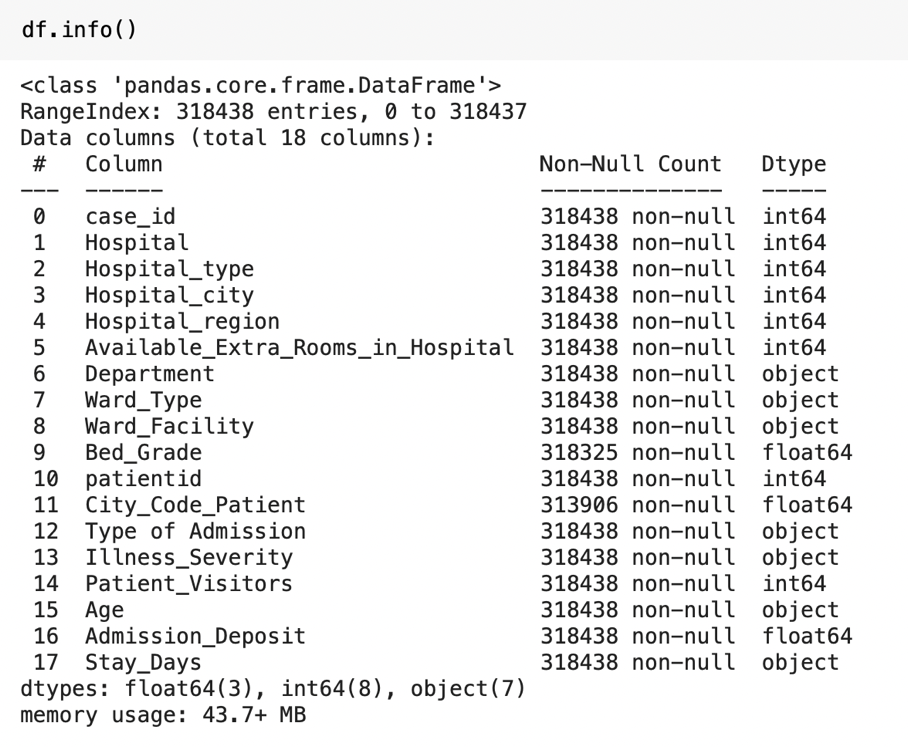




(Figure1)

The datahad the following columns:

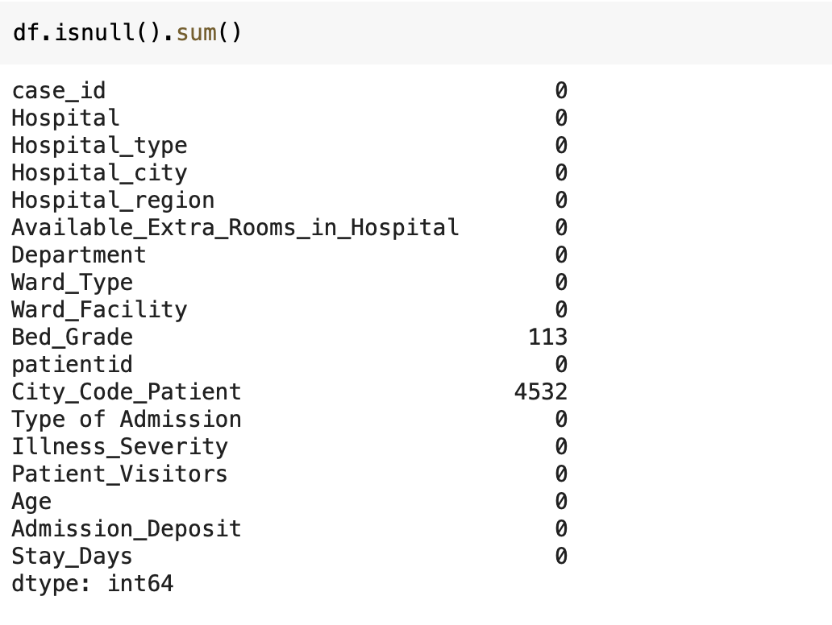
* Case id: Case ID registered in Hospital
* Hospital: Unique code for the Hospital
* Hospital type: The hospitals types include clinics, emergency & Other Outpatient Care Centres etc. which are numbered 1-7
* Hospital city: The city in which hospital lies
* Hospital region: Region Code of the Hospital
* Available Extra rooms in Hospital: Extra rooms available in the hospital
* Department: Department overlooking the case { Radiology, 'anaesthesia' 'gynaecology' 'TB & Chest disease' 'surgery'}
* Ward Type: Code for the Ward type{'R' 'S' 'Q' 'P' 'T' 'U'}
* Ward Facility: Code for the Ward Facility {'F' 'E' 'D' 'B' 'A' 'C'}
* Bed Grade: Condition of Bed in the Ward
* Patient id: Unique Patient Id
* City Code Patient : City Code for the patient
* Type of Admission :Admission Type registered by the Hospital | {'Emergency' 'Trauma' 'Urgent'}
* Illness Severity: term used to characterize the impact that a disease {'Extreme' 'Moderate' 'Minor'}
* Patient Visitors: No. of visitors visiting the patient in due course of treatment
* Age : Age category of patient {'51-60' '71-80' '31-40' '41-50' '81-90' '61-70' '21-30' '11-20' '0-10' '91-100'}
* Admission Deposit : Deposit paid to the hospital at the time of admission
* Stay Days: Number of days for which the patient was under the treatment in hospital. {'0-10' '41-50' '31-40' '11-20' '51-60' '21-30' '71-80' 'More than 100 Days' '81-90' '61-70' '91-100'}
* Info



(Figure 2)

There are 318438 instances in the dataset, which means that it is fairly sufficient as per by Machine Learning standards. 11 of the attributes are numerical and rest all have categorical data.

* Null Value



(Figure 3)

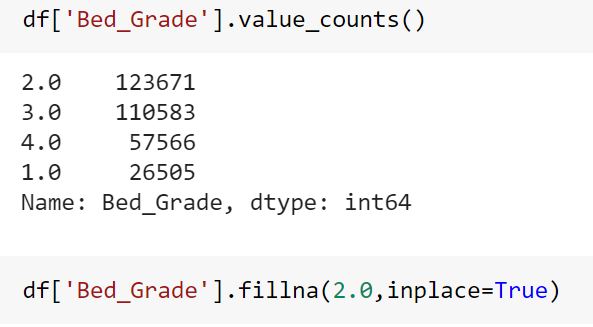
Bed Grade and City\_Code\_Patient features have null values. These columns will be taken into consideration and processed accordingly in further stage.

**3.2.** **Data Preprocessing**

Missing values in a dataset can give misleading results and so are required to be treated.

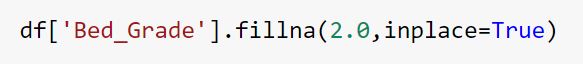
3.2.1. Treating null values in Bed Grade

* Bed\_Grade with value 2 counted to be maximum



(Figure 4)

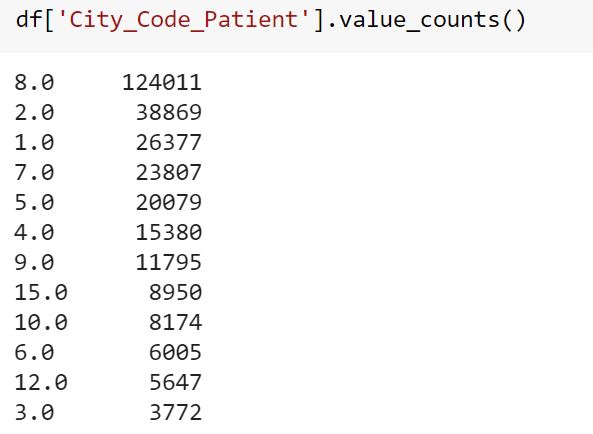
* Missing 113 values were filled by 2.0 Bed\_Grade

****

(Figure 5)

3.2.2 Treating null values in City Code Patient

* City Code Patient with value 8 counted to be maximum



(Figure 6)

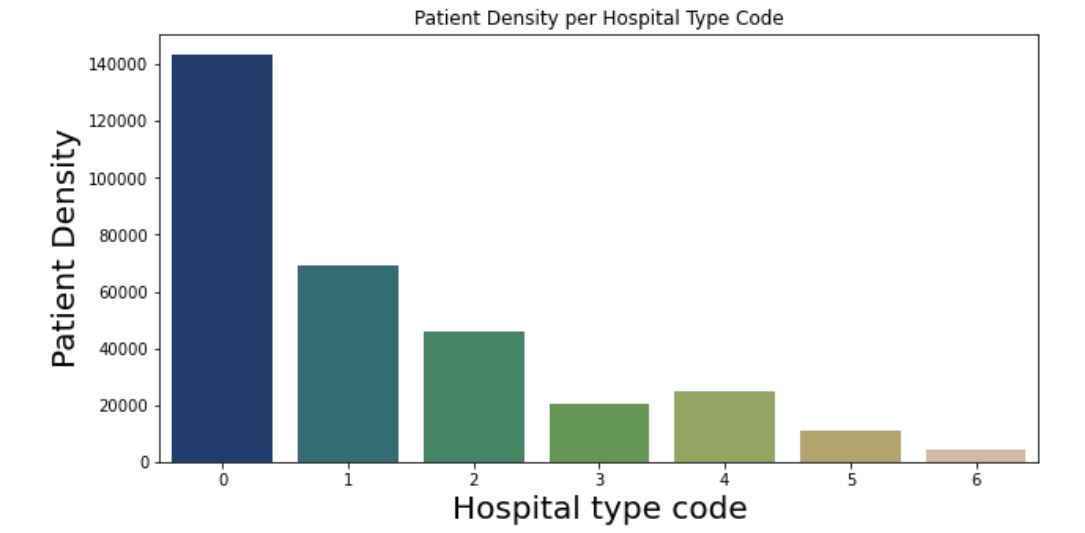
* Missing 4532 values were filled by 8.0 in City\_Code\_Patient



(Figure 7)

**3.3. Data Visualization**

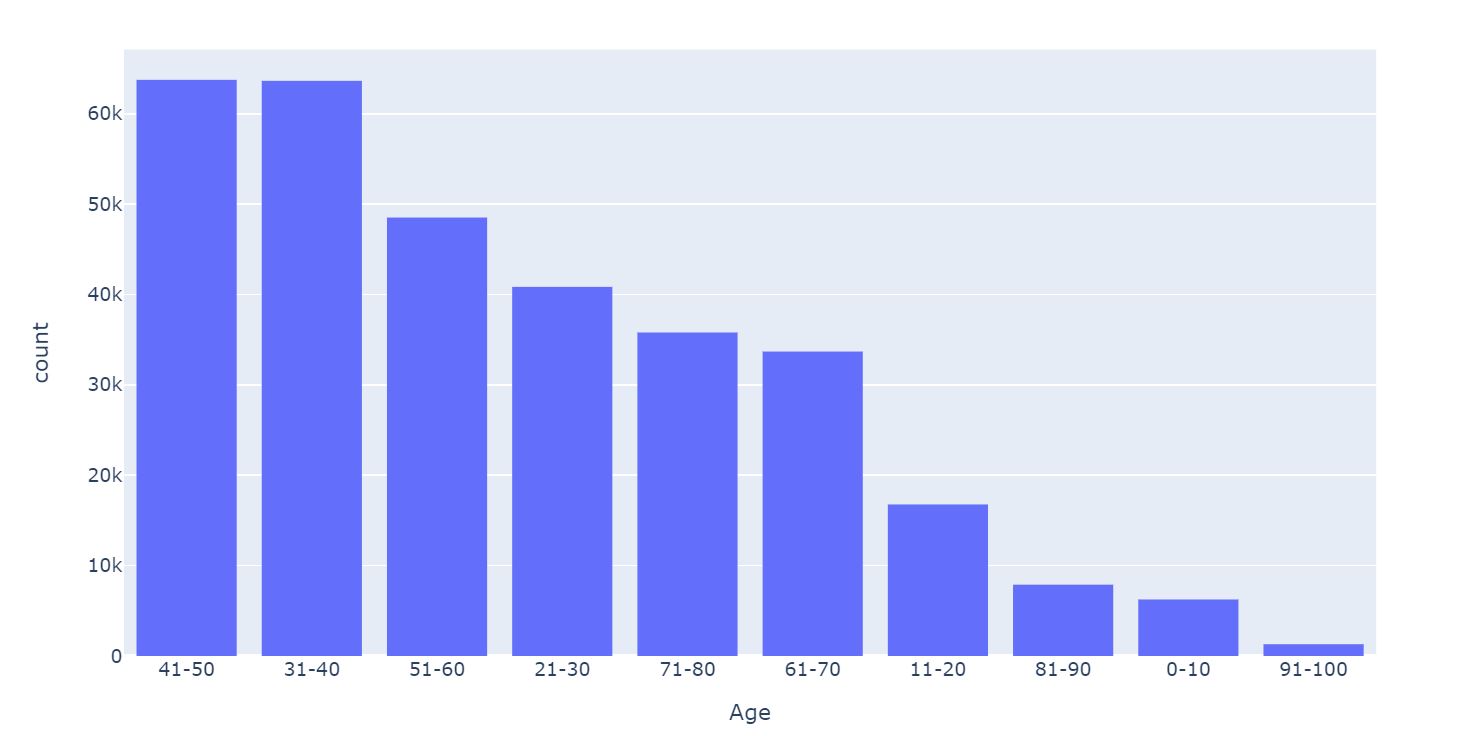
The data was analysed through visual representation using Matplotlib ,Seaborn & Plotly Express.

* Bar Graph showing hospital type with patient density in hospital

(Figure 8)

Hospital with type code'0' has highest number of patients thus, less beds/rooms left available. Hospital with type code'6' has lowest number of patients thus, more beds/rooms left Bar Graph showing count of patients admitted in hospitals.

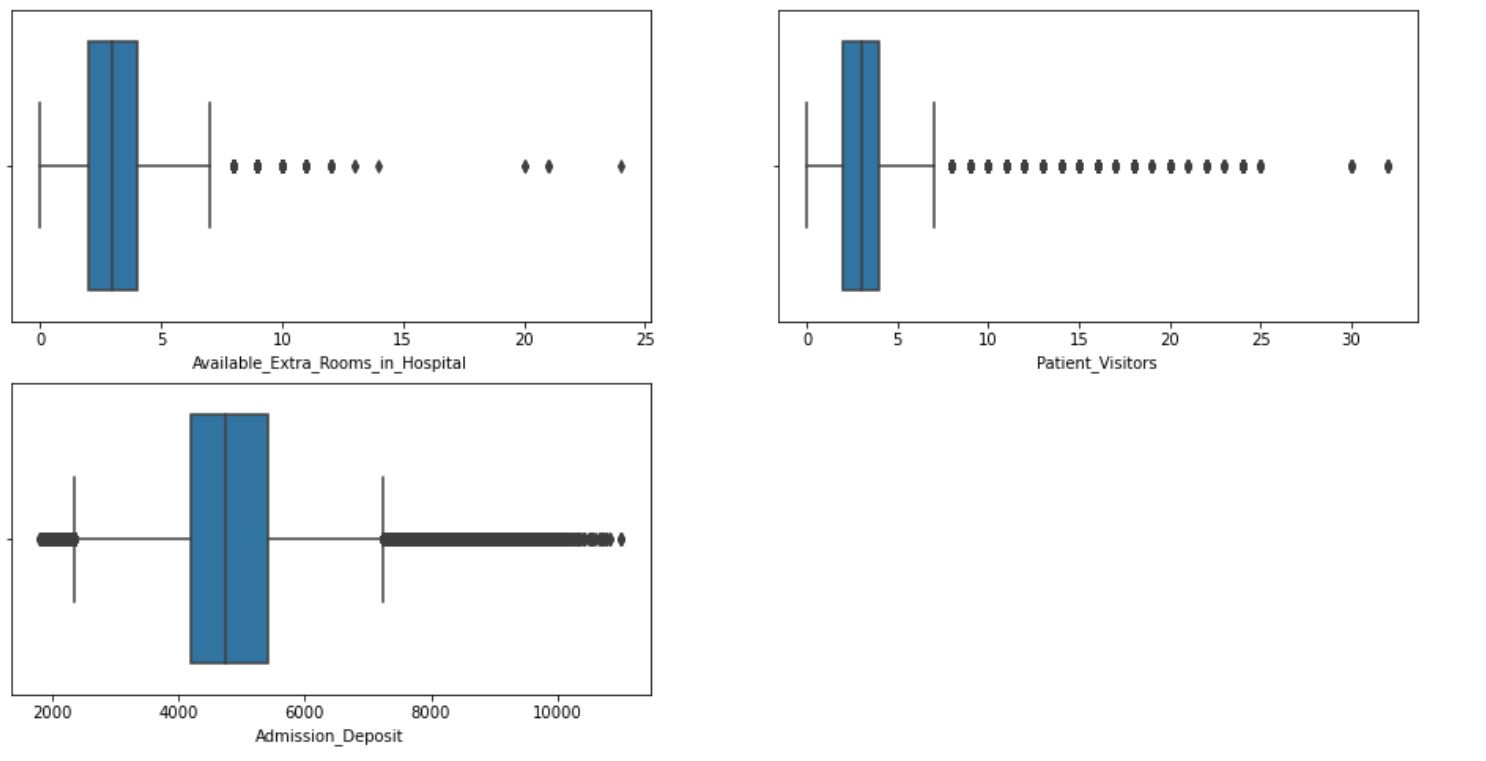
* Bar Graph showing Age and count of people admitted



(Figure 9)

Mostly patients admitted are between the age of 30-50. The other age group are been admitted comparatively less. The admission in hospital declines in later years especially after 71 year of age.

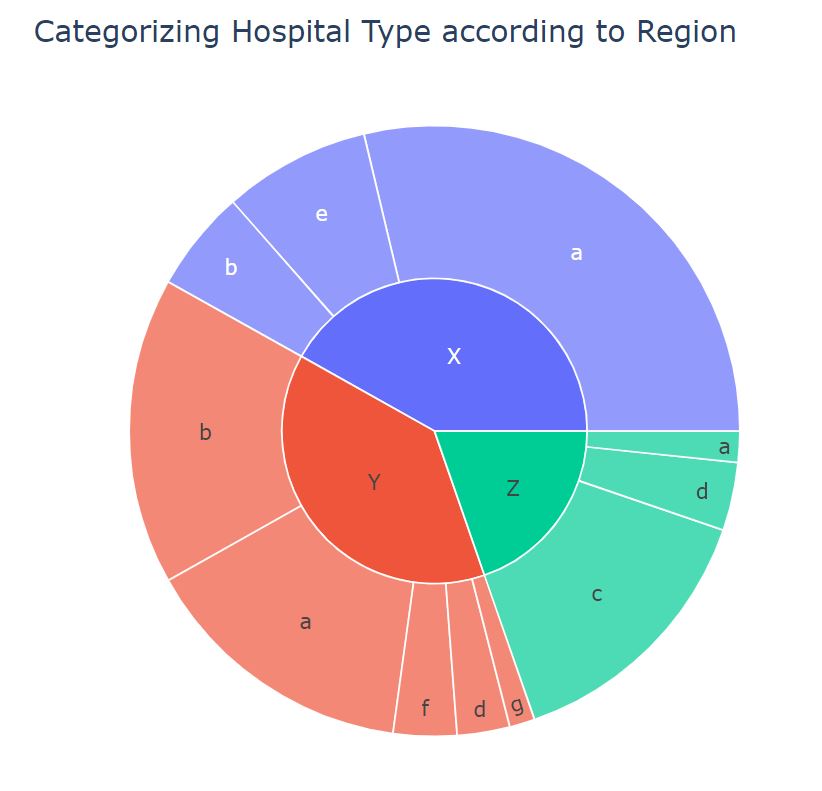
* Boxplot showing Available\_Extra\_Rooms\_in\_Hospitals, Patient\_Visitors and Admission\_Deposit



(Figure 10)

Available\_Extra\_Rooms\_in\_Hospitals, Patient\_Visitors and Admission\_Deposit show significant number of outliers.

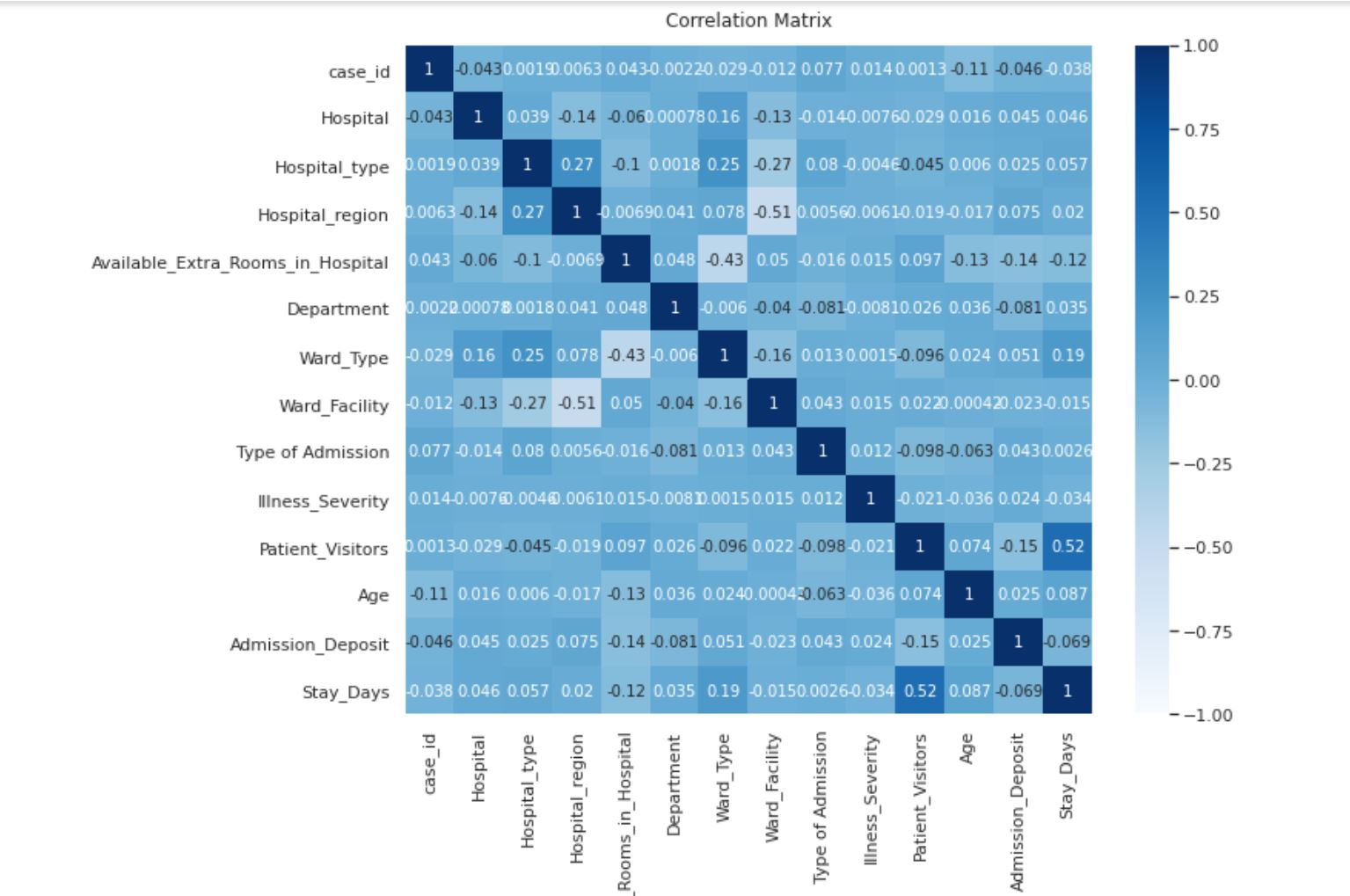
* The below figure shows density of people in different hospital type according to 3 different region i.e. X,Y &Z.



(Figure 11)

This depicts that patient density is very close within X and Y region in the X region the patients are better distributed between different hospital codes than regions Y and Z.

* A correlation plot of all the attributes of data

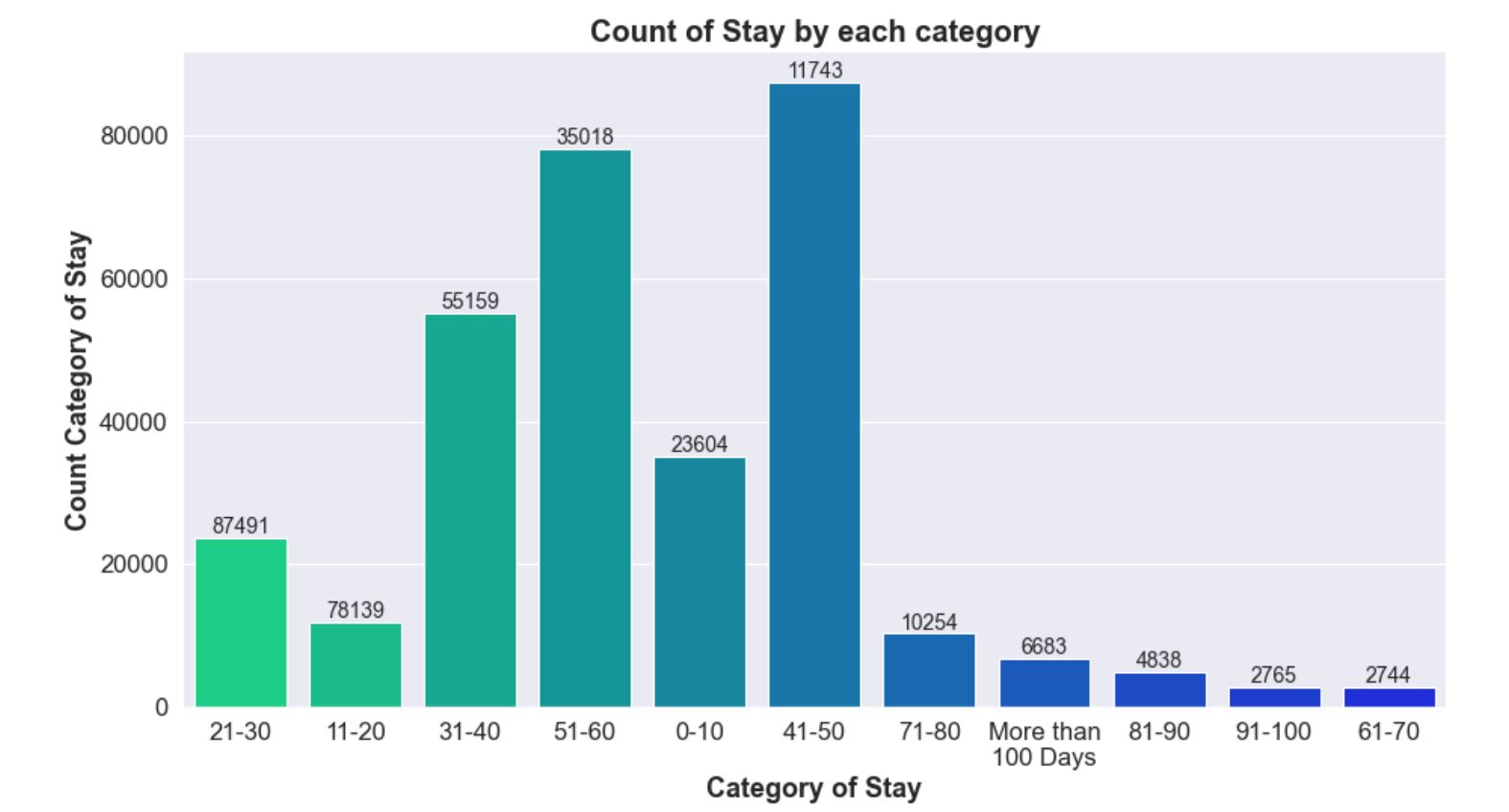


(Figure 12)

Correlation matrix was used to analyse the relationship between the features. Correlation between variables are less than expected. The least among the given ones need to be removed. Columns: 'patient ID', 'Hospital city', 'Bed Grade' & 'City Code Patient' don't influence stay days. Hence these columns will be deleted.

**Stay Days Analysis**

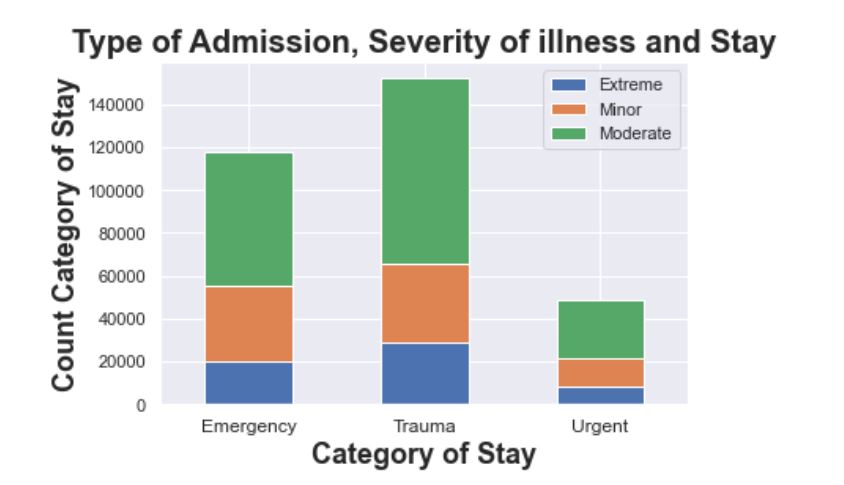
* Bar graph here represents count of patients admitted in different categories of stay days



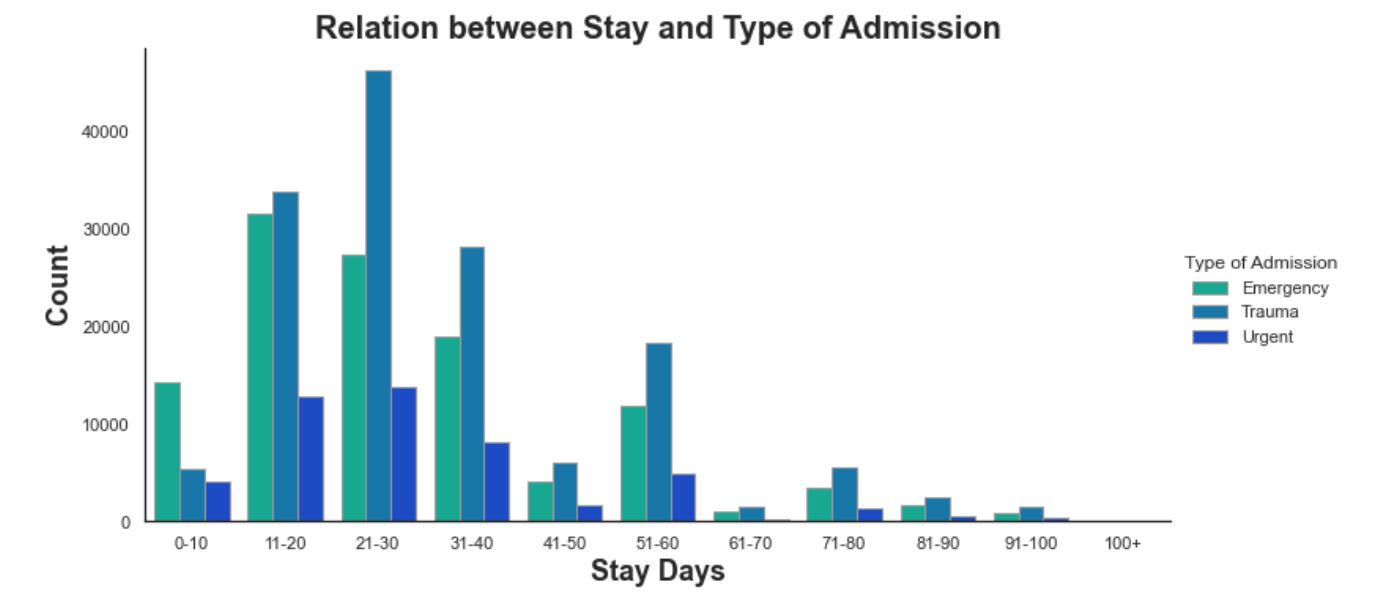
(Figure 13)

Most number of patients stay for less than 50 days. However as the number of day’s increases 50 a sharp decline of number of patients has been observed.

* Both the graphs depict show count category of stay and type of admission



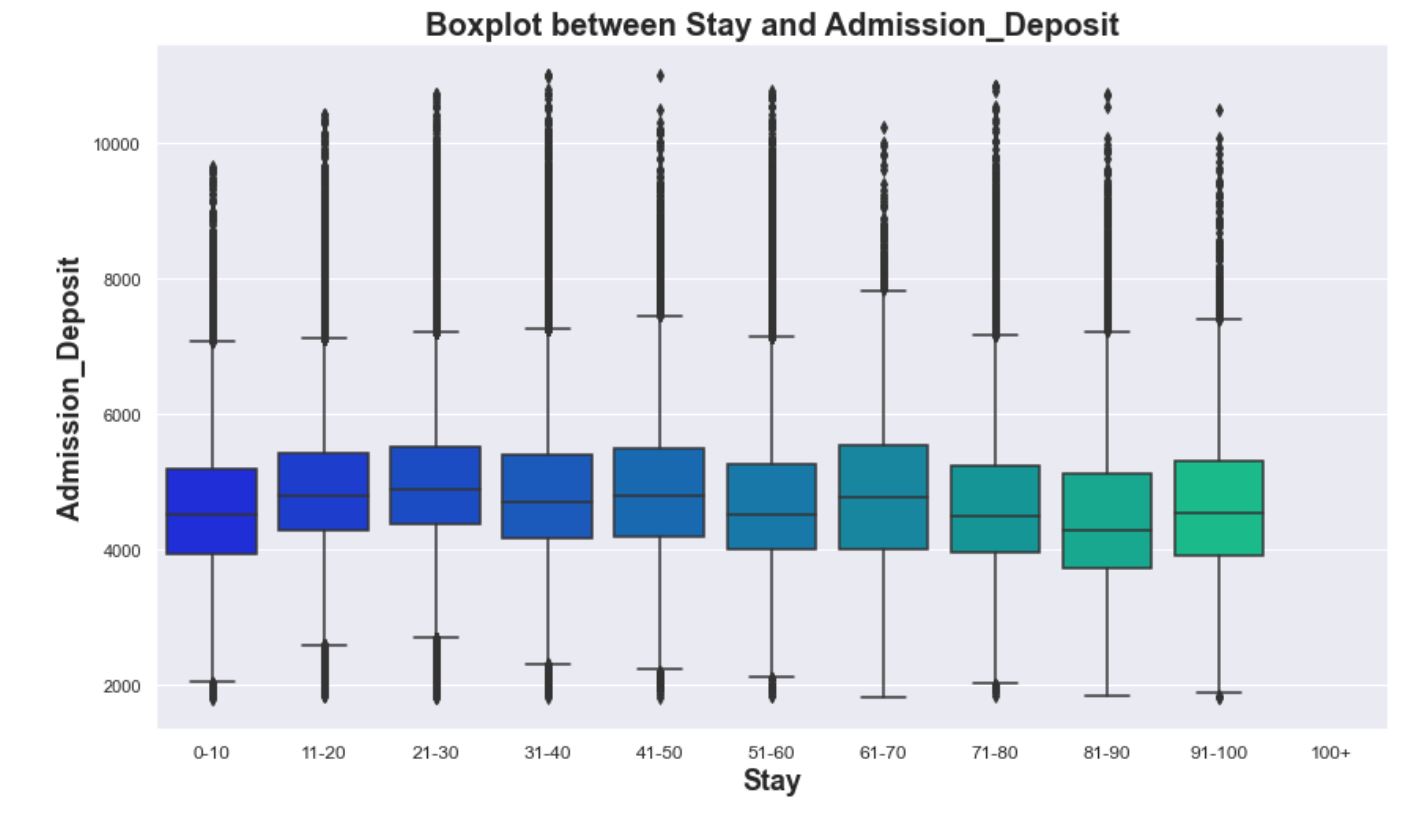
(Figure 14)



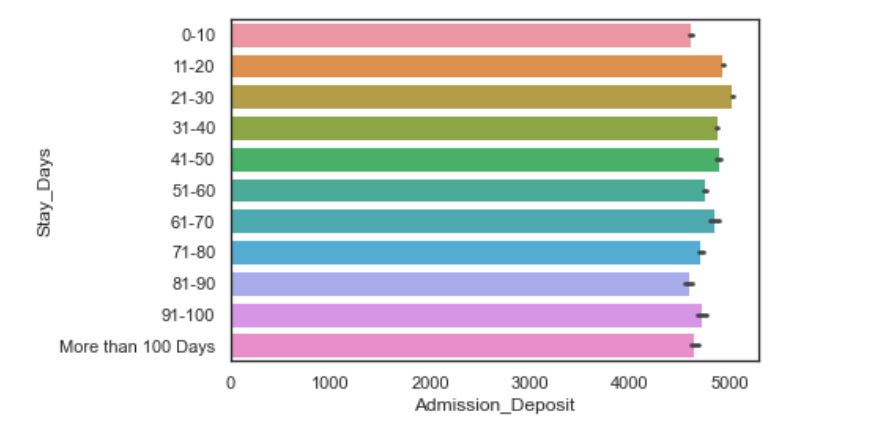
(Figure 15)

Patients mainly suffer from moderate level of illness with admission type Trauma. Showing greater requirement of equipment’s related to 'moderate ‘Illness and admission type 'Trauma'.

* Stay Days and amount Admission\_deposit are depicted through 2 of the plots below i.e. Boxplot and bar graph



(Figure 16)

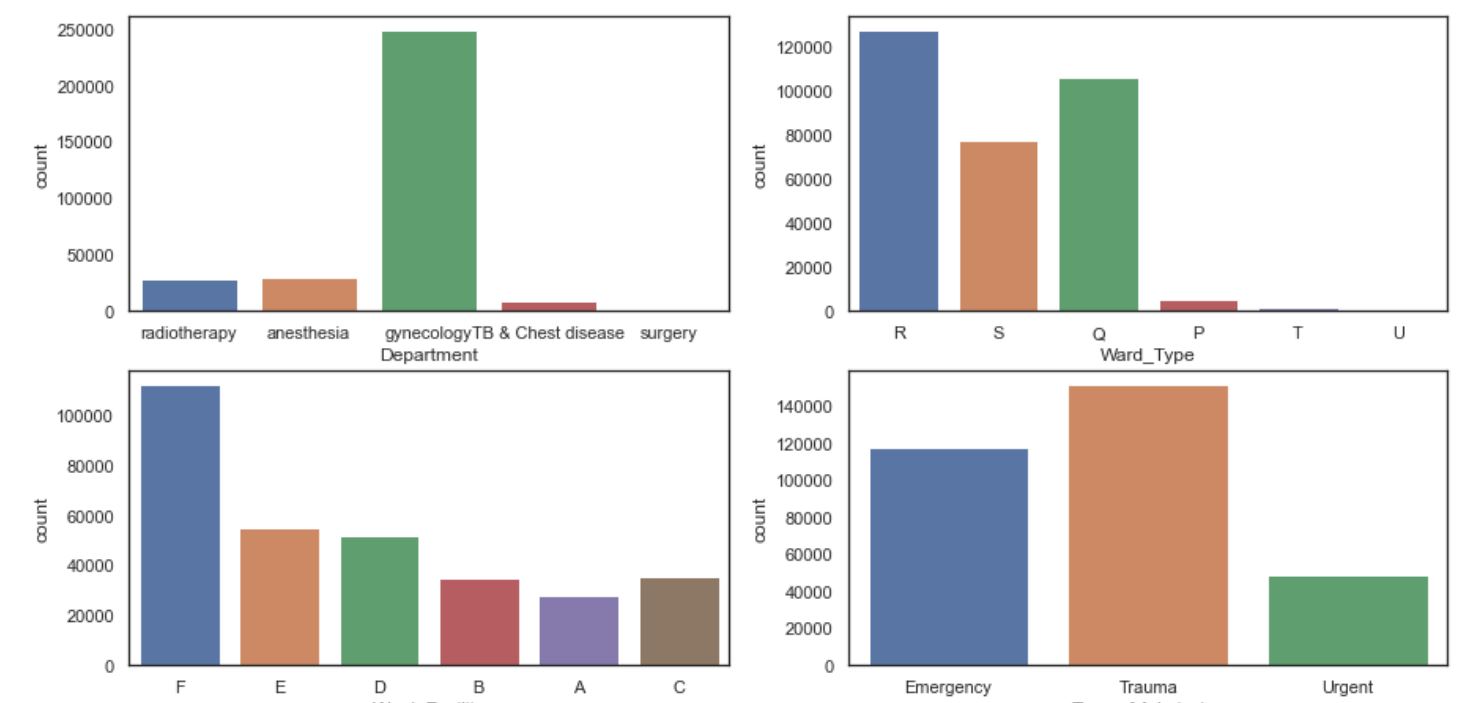


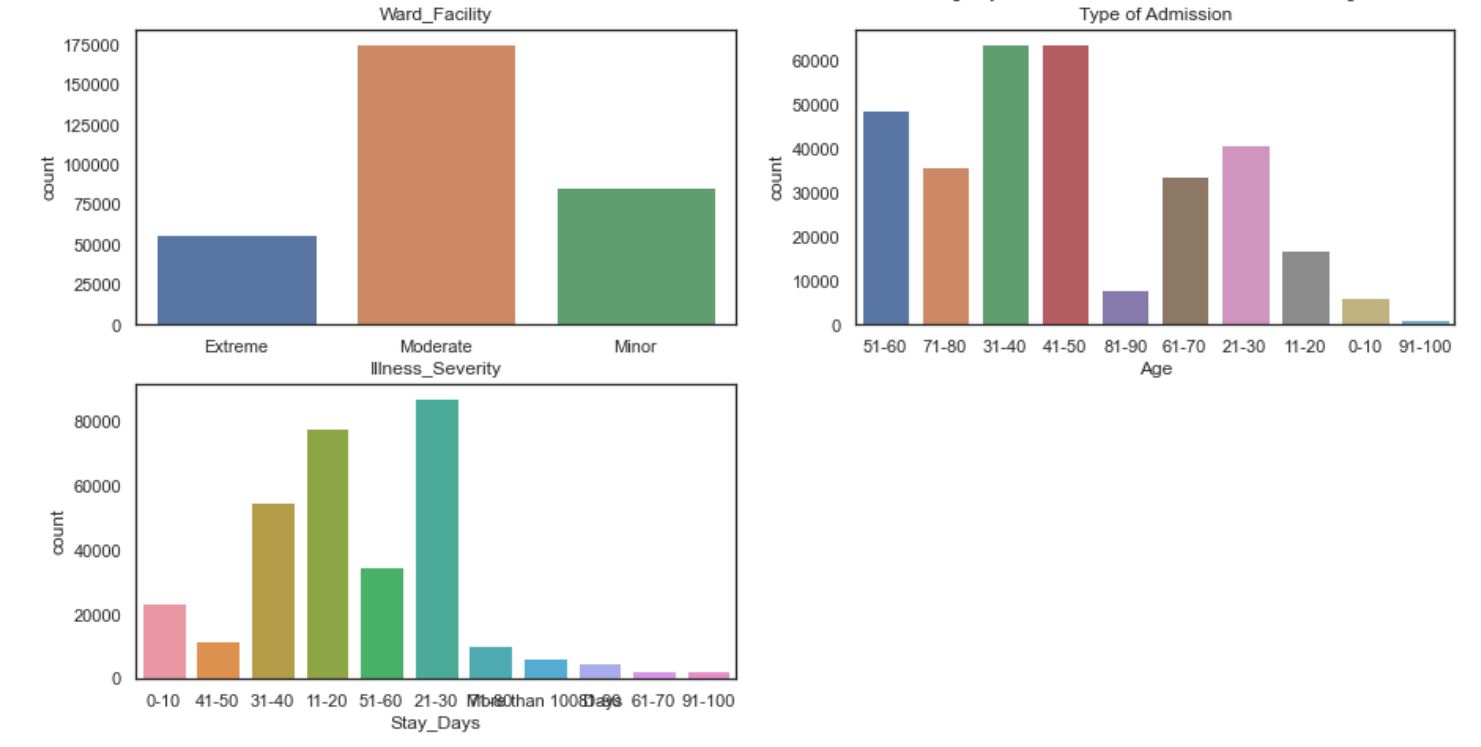
(Figure 17)

The above graphs shows that when the Admission Deposit tend to have relatively low stay days. This may due to patients being discharged early.

**Analysis of Categorical & Numerical Data**

* Subplot was created to analyse the categorical variables i.e. containing Department, Ward\_Type, Ward\_Facility, Type of admission, Illness severity, Age and stay days





(Figure 19)

It was observed that number of patients admitted are significantly inclined towards a particular sub category:

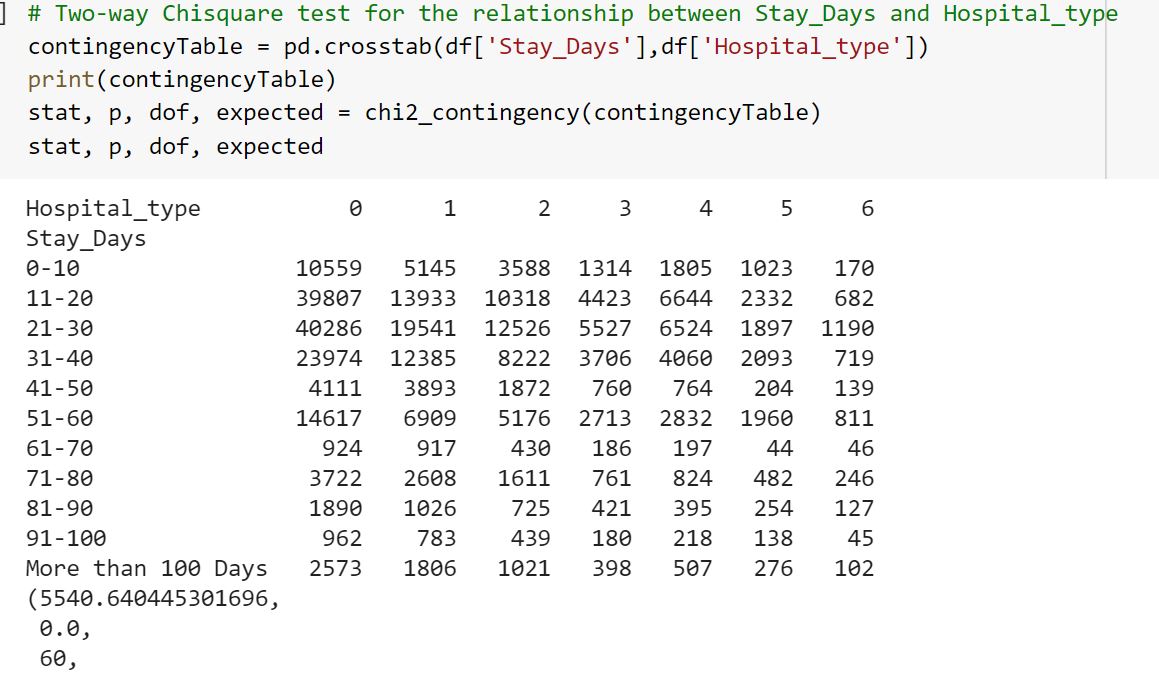
* Department: Gynecology
* Ward Type: R
* Ward Facility: F
* Type of admission: Trauma
* Illness severity: Moderate
* Age: 41-50
* Stay Days: 41:50

**3.4. Determining relationship between Categorical Variables with Stay Days**

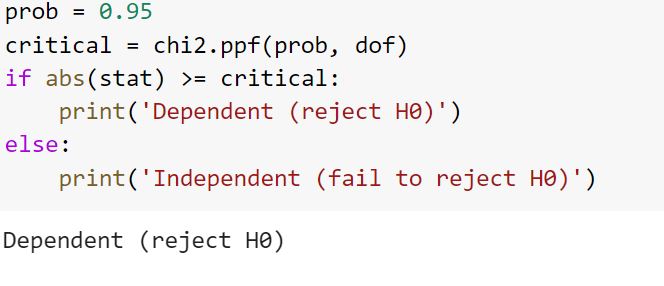
Since Stay\_Days is the target variable it is important to understand relationship between categorical variables present in the data.

Chi Square testing is statistical technique used to determine relationship between categorical variables:

* Two-way Chi Square test for the relationship between Stay\_Days and Hospital\_type

****

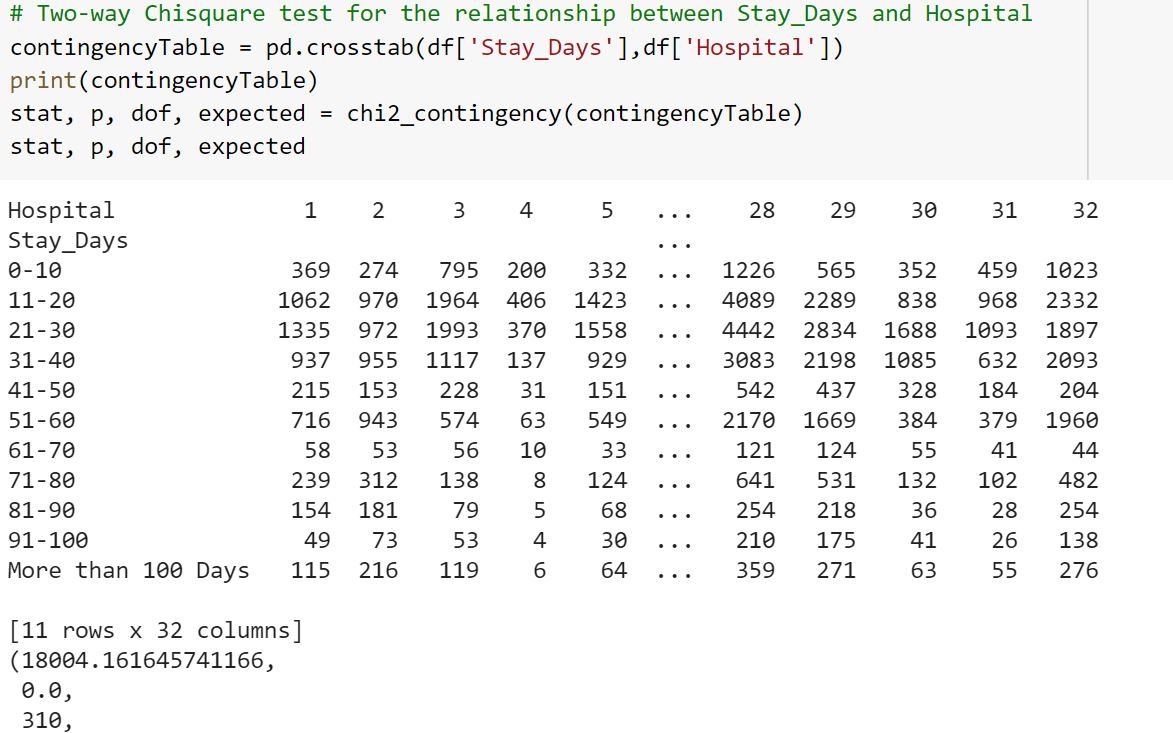
(Figure 20)

****

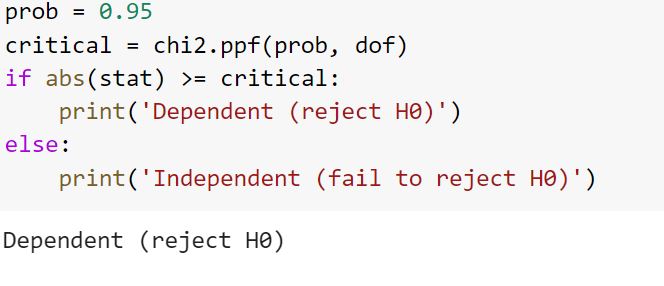
(Figure 21)

Hospital\_type is statistically dependent on stay days by doing two way Chi Square Testing

* Two-way Chi Square test for the relationship between Stay\_Days and Hospital

****

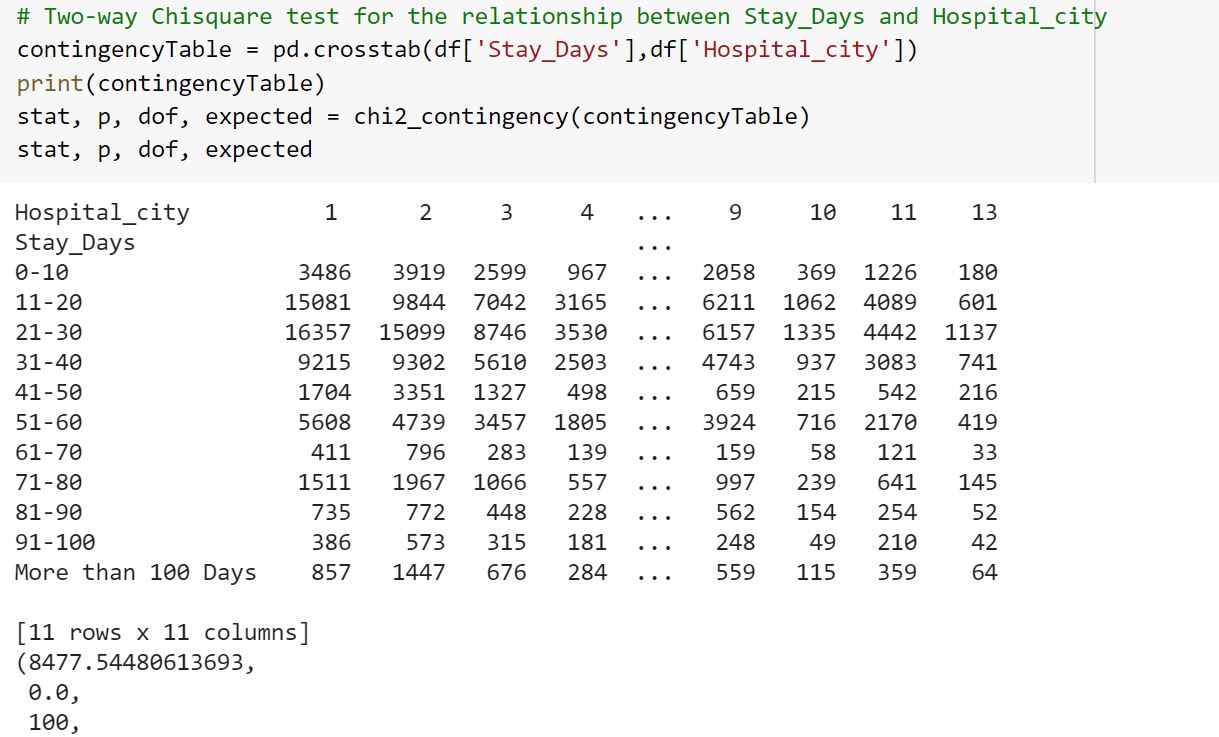
(Figure 22)

****

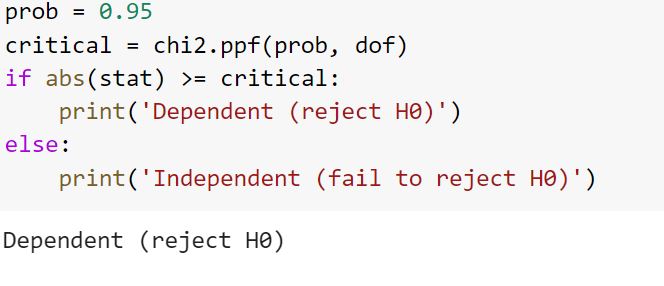
(Figure 23)

Hospital is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and Hospital

****

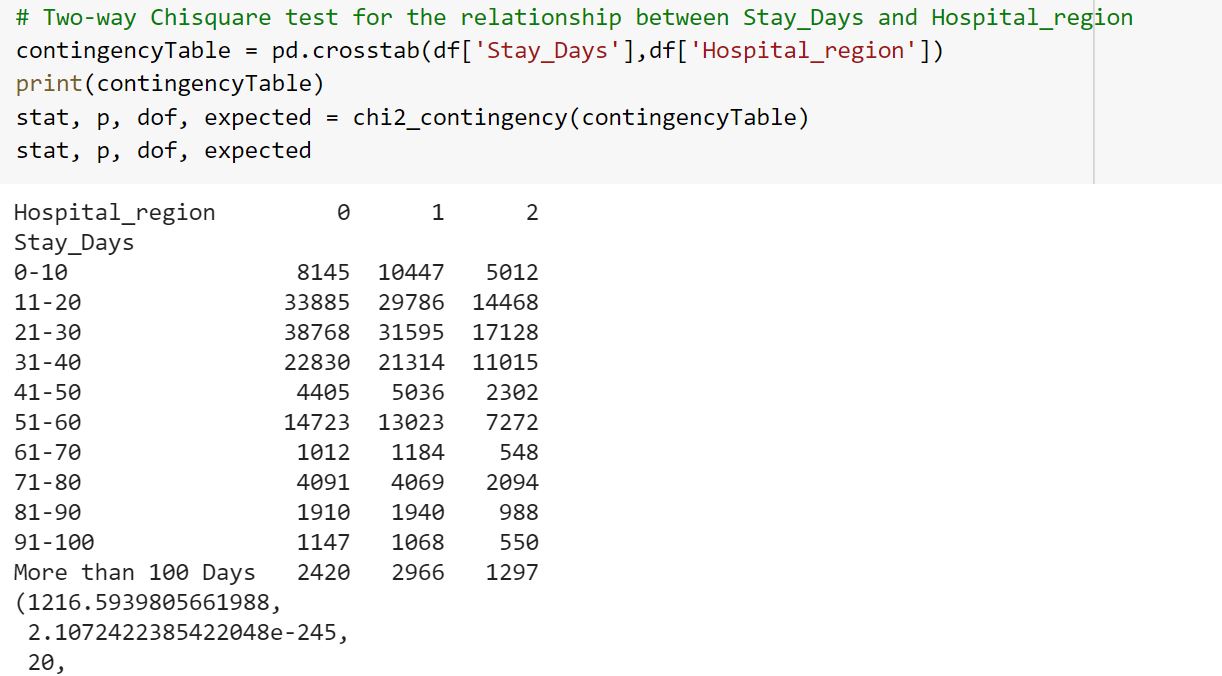
(Figure 24)

****

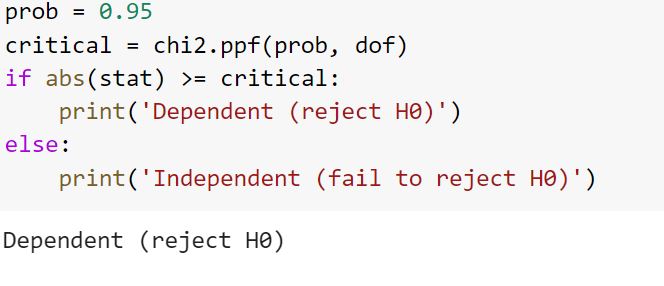
(Figure 25)

Hospital\_city is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and Hospital\_Region

****

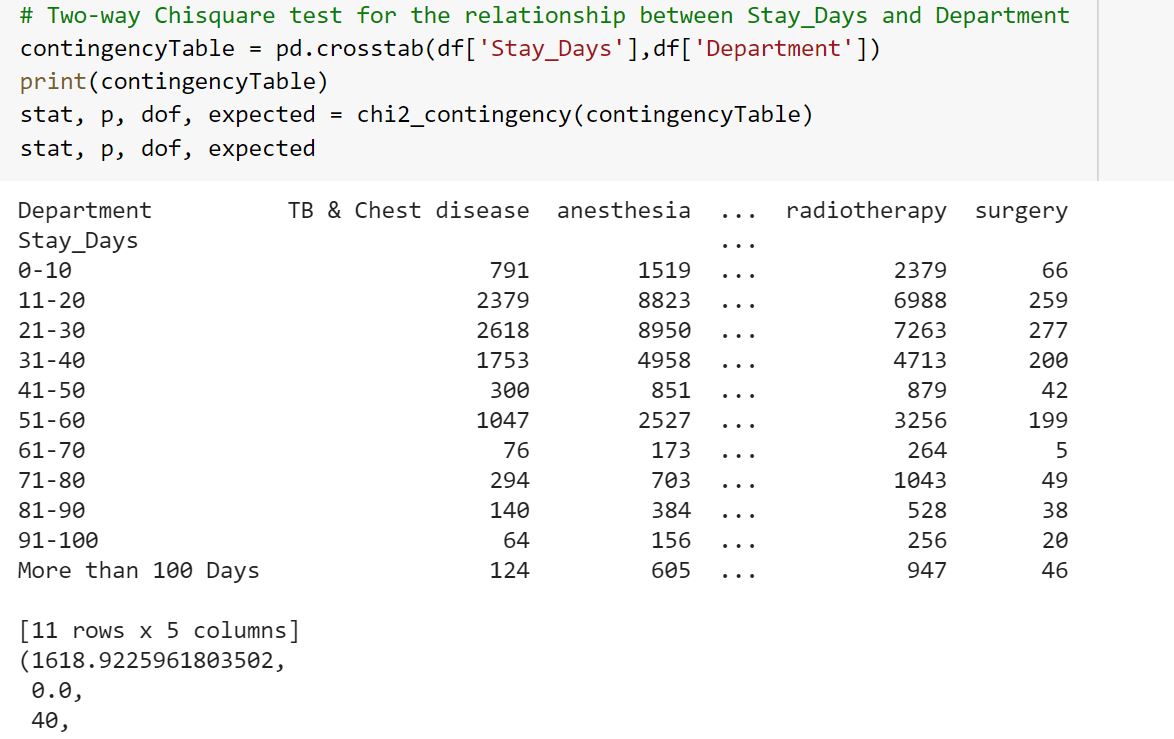
(Figure 26)

****

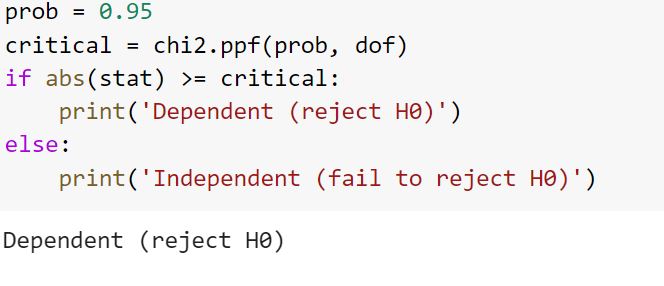
(Figure 27)

Hospital\_region is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and Department

****

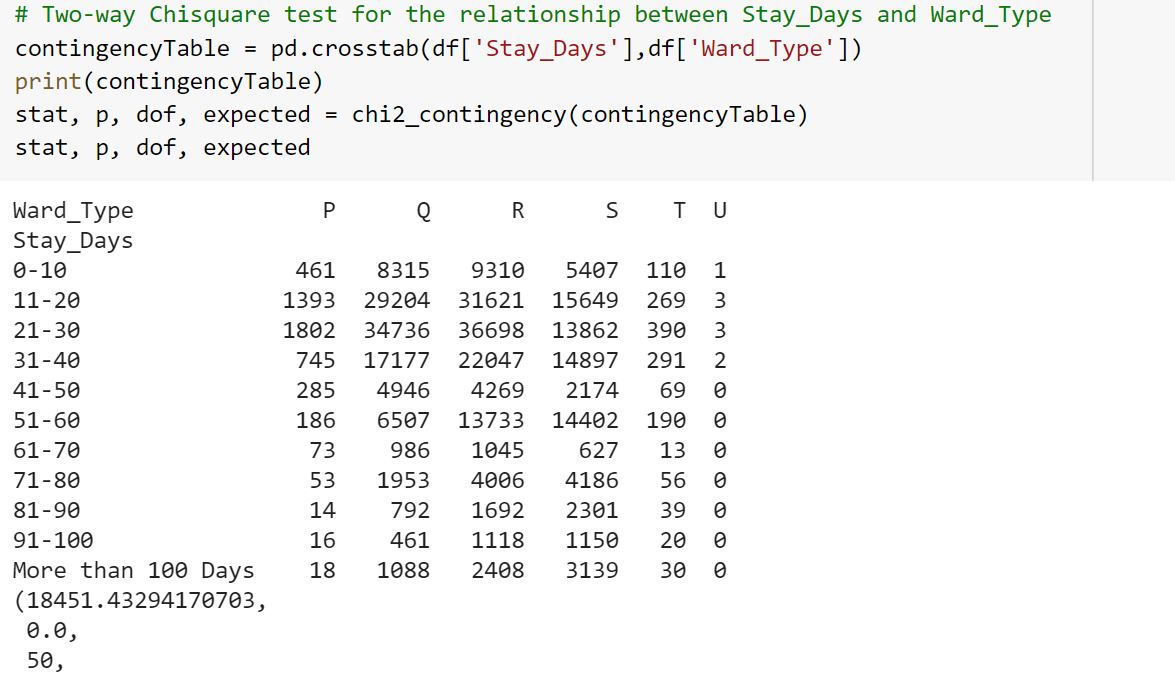
(Figure 28)

****

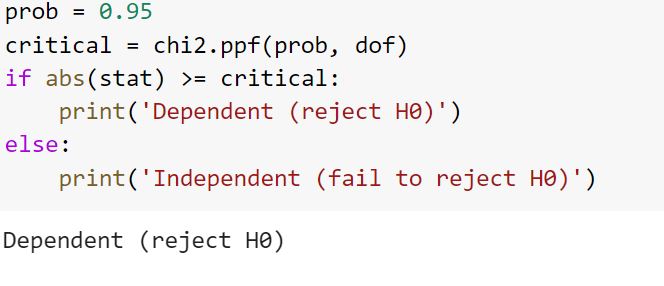
(Figure 29)

Department is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and Ward\_Type

****

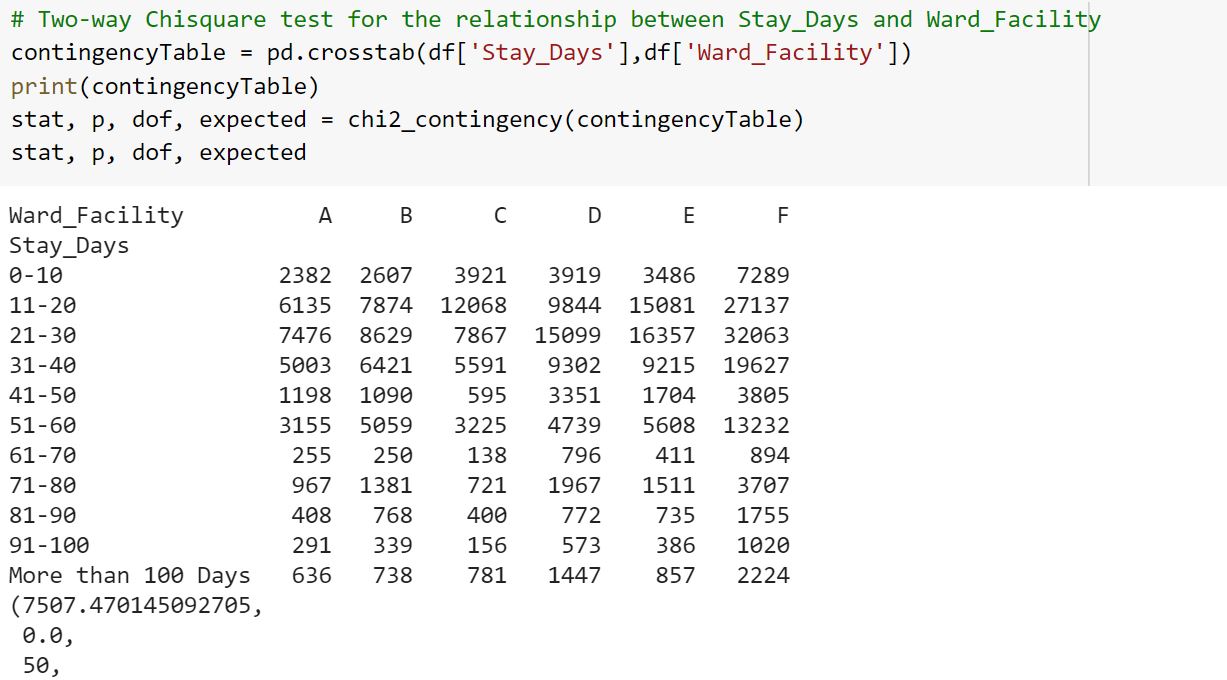
(Figure 30)

****

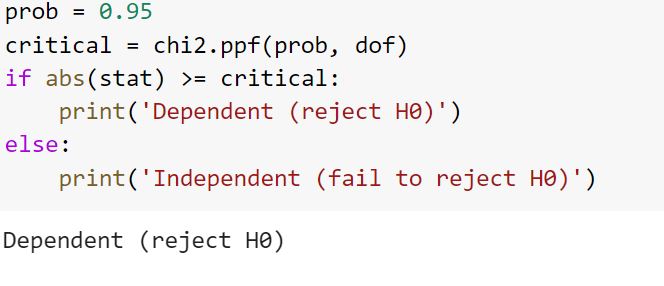
(Figure 31)

Ward\_Type is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and Ward\_Facility

****

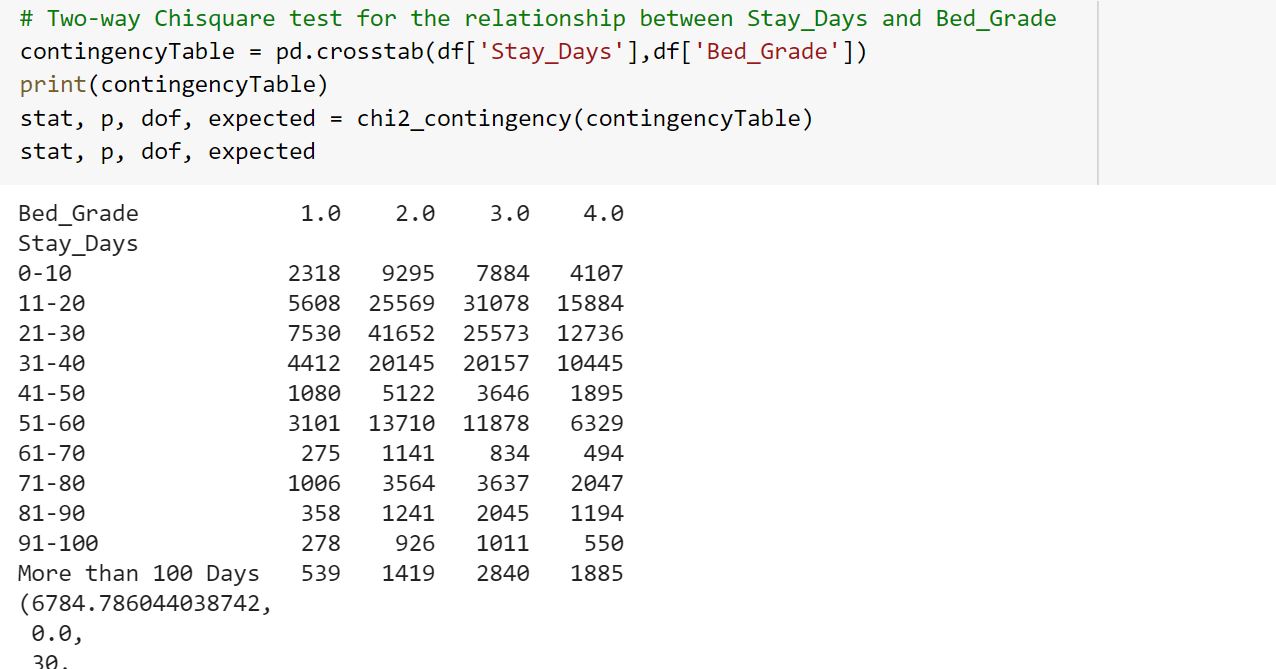
(Figure 33)

****

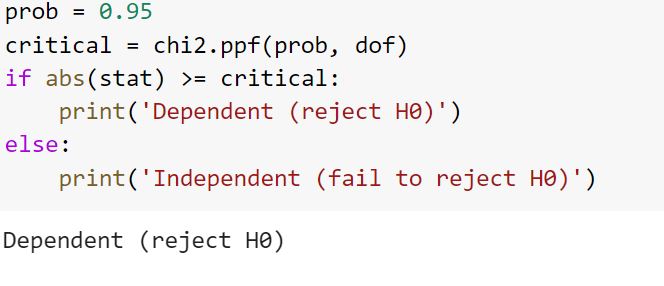
(Figure 34)

Ward\_Facility is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and Bed\_Grade

****

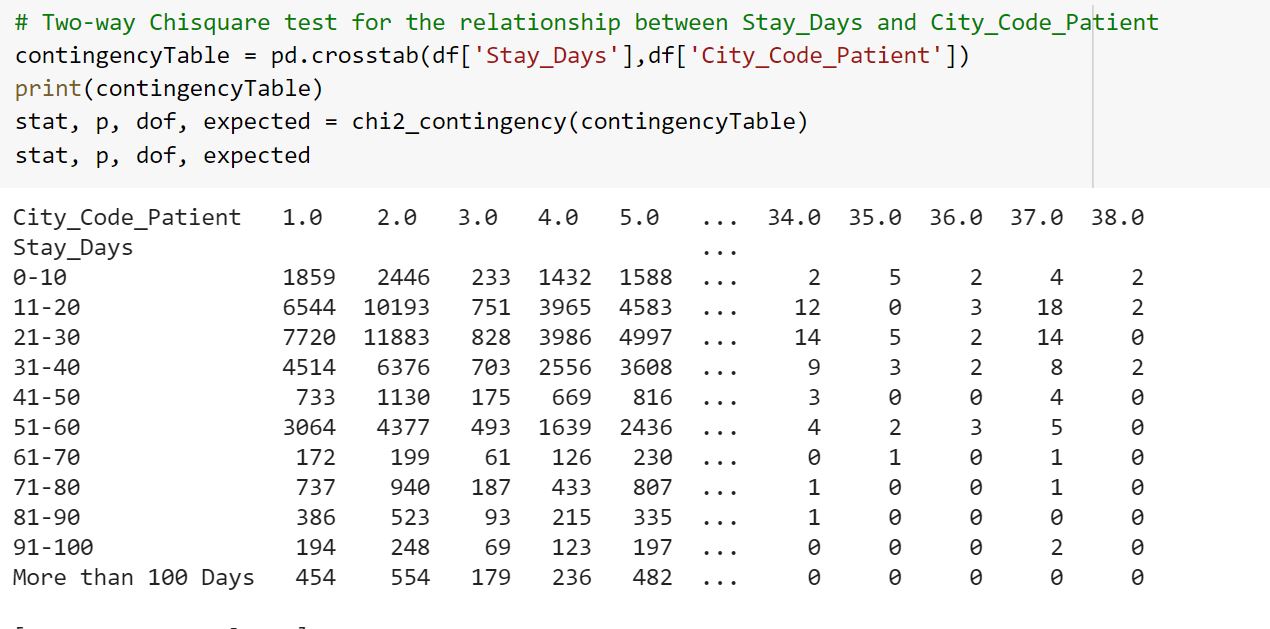
(Figure 35)

****

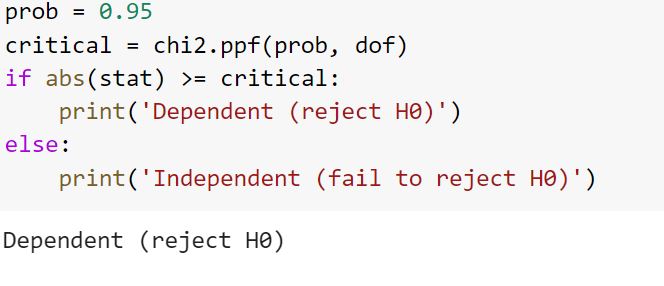
(Figure 36)

Bed Grade is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and City\_Code\_Patient

****

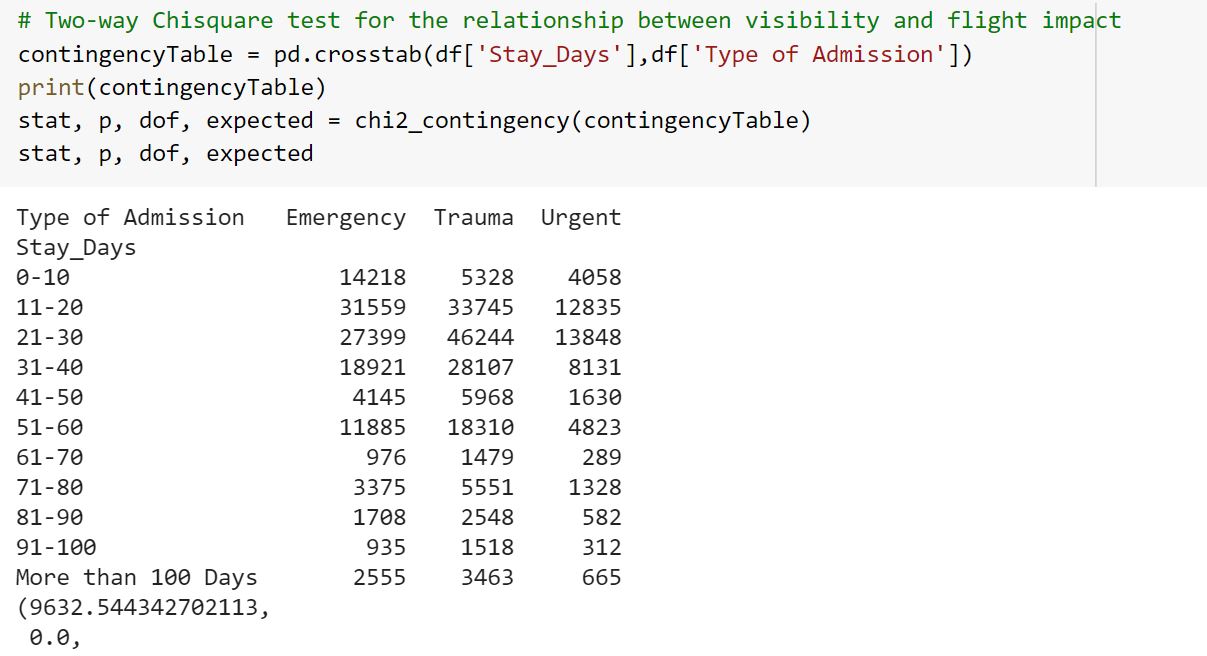
(Figure 37)

****

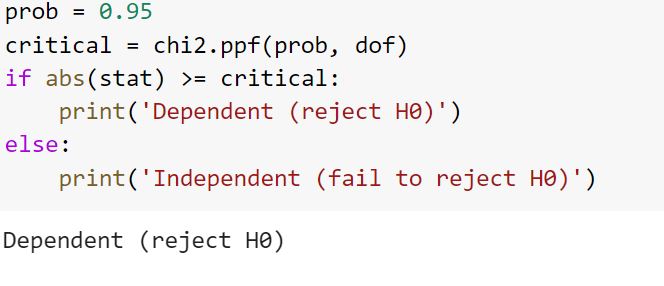
(Figure 38)

City\_code\_patient is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and City\_Code\_Patient

****

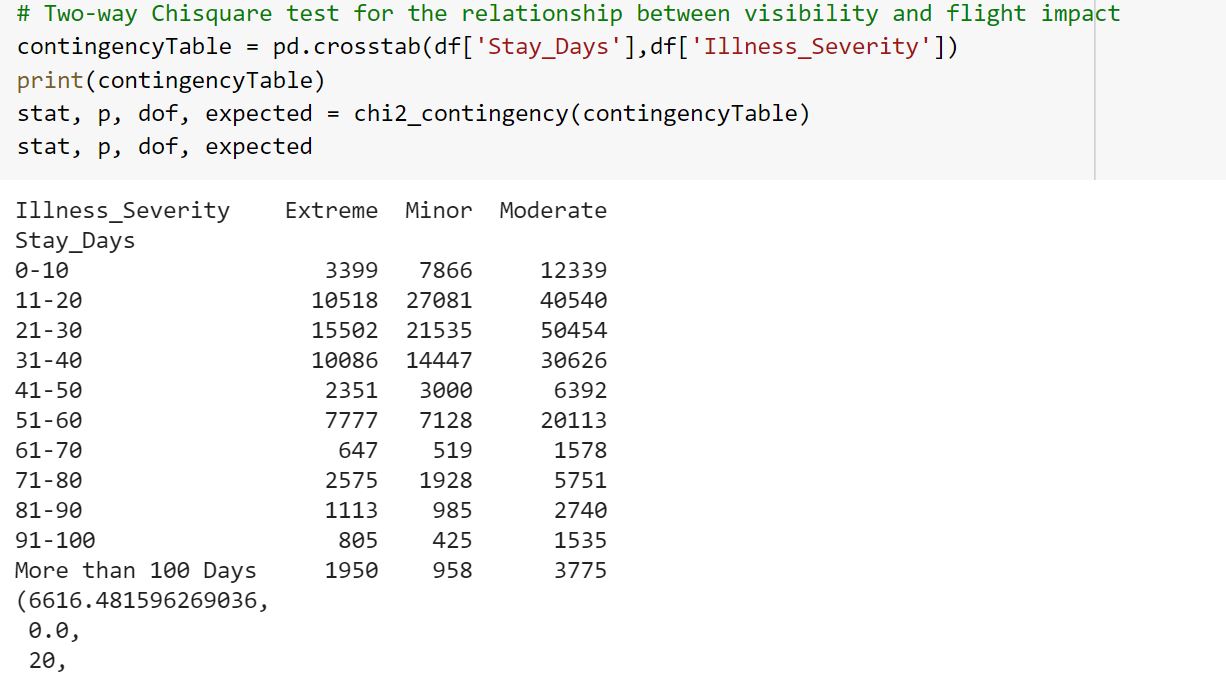
(Figure 39)

****

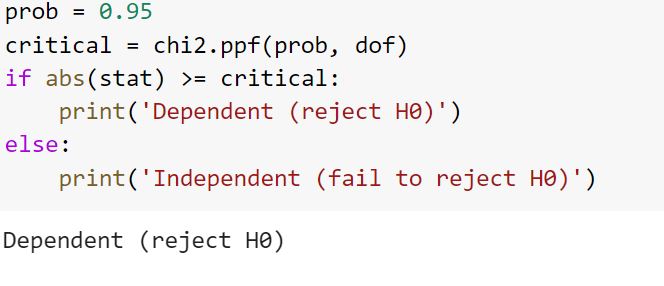
(Figure 40)

City\_code\_patient is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and Illness Severity

****

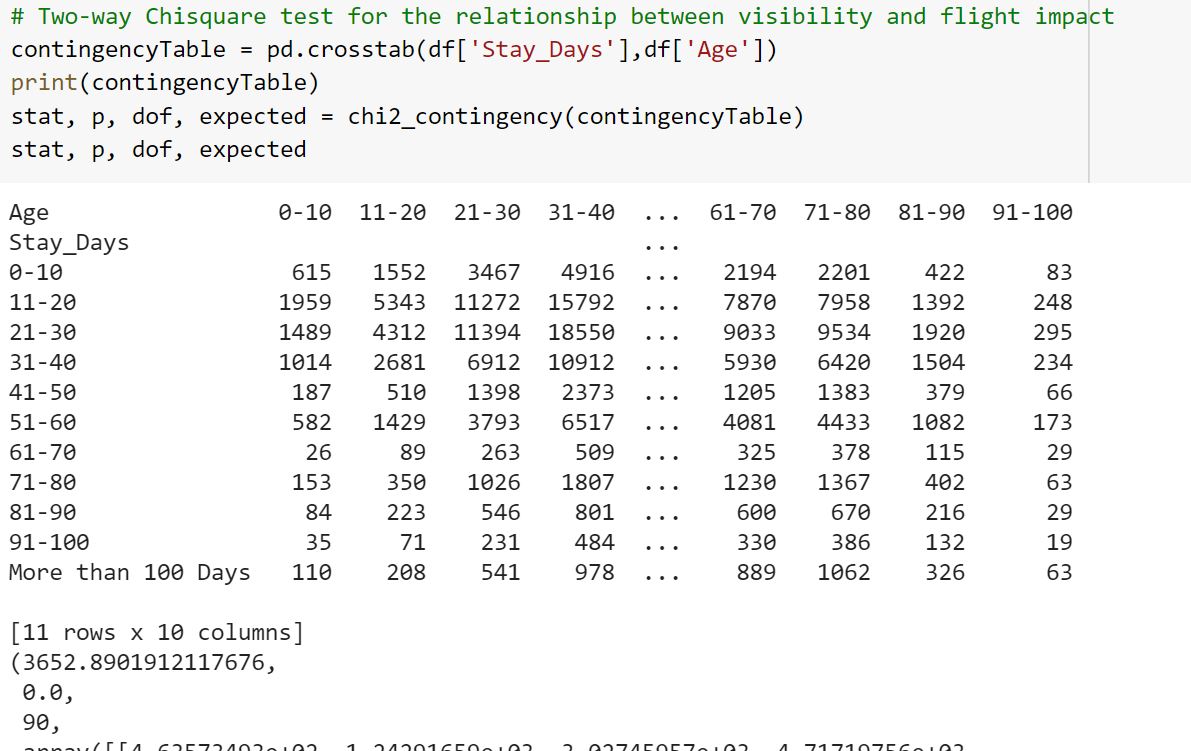
(Figure 41)

****

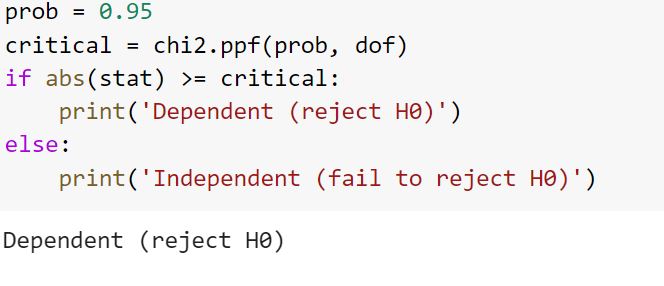
(Figure 42)

Illness Severity is statistically dependent on stay days by doing two way Chi Square Testing.

* Two-way Chi Square test for the relationship between Stay\_Days and Age

****

(Figure 43)

****

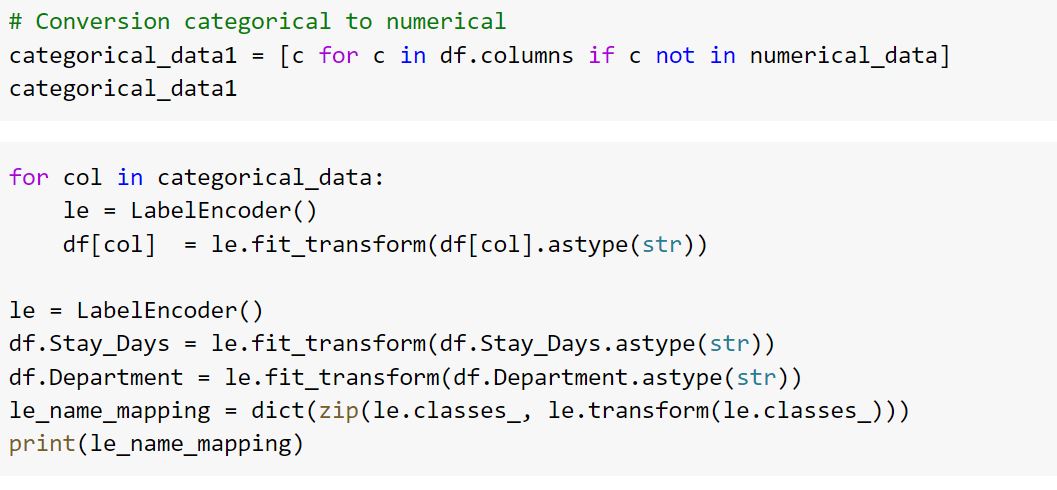
(Figure 44)

Age is statistically dependent on stay days by doing two way Chi Square Testing.

All the Categorical variables came to be statistically dependent on stay days by doing Chi Square Testing of all the categorical variables.

**3.5. Data Engineering**

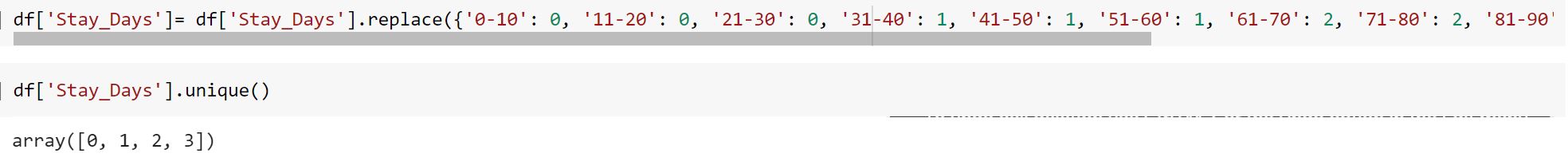
**3.5.1. Label Encoding**

****

(Figure 45)

Since the data contained it required to be converted into numerical. Label Encoder from Scikit Learn library encoded categorical columns to numerical column.

**3.5.2 Splitting into classes**



(Figure 46)

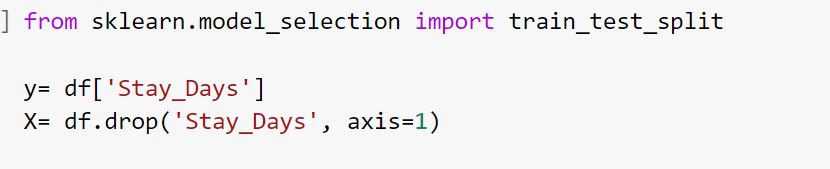
Stays was initially divided into 10 classes but on applying the model the accuracy came out to be very low. So Stay\_Days was converted into into 4 classes and allotted the following thereon:

* 0-30: 0
* 31-60: 1
* 61-90: 2
* 90 and above: 3

**3.7. Model Building**

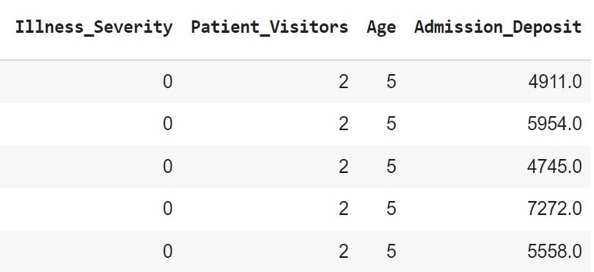
**3.7.1. Splitting the Data**

* The combined data was separated to X and Y where stay days was the target variable that are to be predicted**.**

****

(Figure 47)

* X has 13 attributes i.e case-id, hospital, hospital type, region, available extra rooms, department, ward facility, admission type, illness severity, patient visitors, age, admission deposit and Y has the attribute “Stay Days” that is to be predicted

****

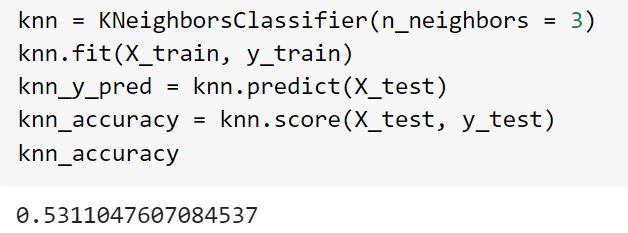
(Figure 48)

* The data was further divided into training and testing data in ratio 80 and 20 respectively. So the train data consists of 254750 data points and test data consists of 63688 data points

**3.7.2. Modelling**

3 ML algorithms to implement a LOS prediction model: KNN, Decision Tree and Random Forest.

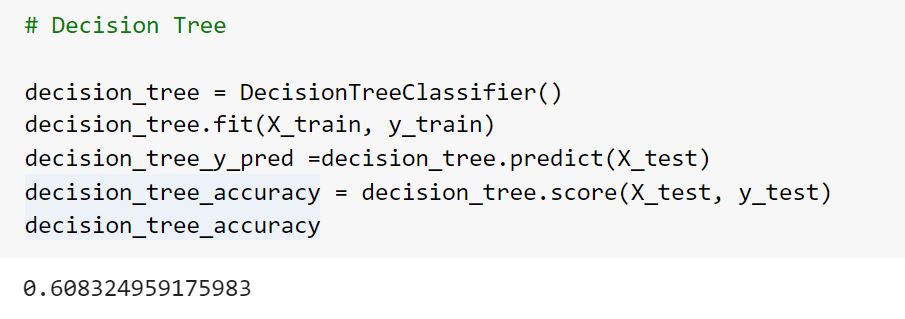
* **KNN Classifier**:



(Figure: 49)

K Neighbors Classifier showed an accuracy of 0.53110 on putting n\_neighbors as 3.

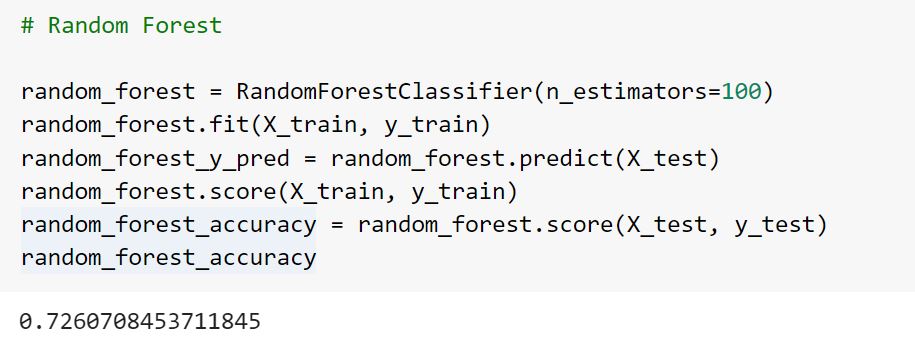
* **Decision Tree:**



(Figure 50)

Decision showed an accuracy of 0.69832 which is higher than K Neighbors but still not satisfying.

* **Random Forest:**

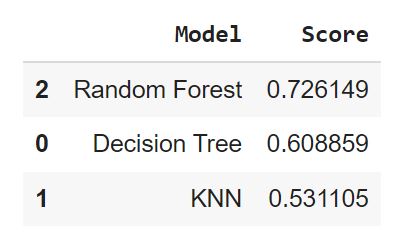


(Figure 51)

Random Forest Classifier showed an accuracy of 0.7260 with n\_estimators as 100 because of the data size of the study which is fairly large.

This level of accuracy is satisfactory in comparison to other models used.

The system was trained on a dataset containing similar information. The accuracy of the used models are:



(Figure 52)

**Chapter 4**

**Results and Discussion**

We tried to determine Length of Stay of Patients. Some features turned out to be of importance while features like city code of patients, hospital etc. turned out to have been less correlated and so were dropped.

Stay Days was divided into smaller number of classes .Categorical data was converted into numerical with the help of label encoder.

In this paper, we investigated the factors impacting the LOS and the most known ML methods.

All the categorical variables are dependent on stay days as per chi-square testing.

Through data visualization it was observed that:

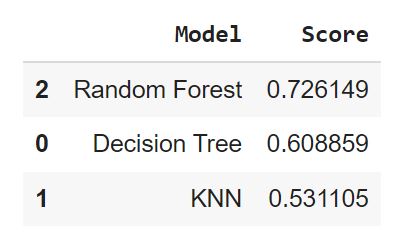
* Patients mainly suffer from moderate level of illness with admission type Trauma.
* The Admission Deposit tend to have relatively low stay days probably because they leave early with higher deposit.
* Also, mostly patients admitted are between the ages of 30-50.
* Though most of the attributes showed low degree of relationship as per the correlation plot. But 'patientid','Hospital\_city','Bed\_Grade' and 'City\_Code\_Patient' showed the least. Since they would show discrepancy in prediction these columns were dropped.

After splitting the data in random state 42 machine learning models were applied.

On using three ML algorithms to implement a LOS prediction model: KNN, Decision Tree and Random Forest.

Initially when the target variable was divided into 10 classes the accuracy level was 0.223 for KNN Classifier, 0.323 for Decision Tree and 0.424 for Random Forest. This level of accuracy showed requirement of more level of data engineering to be done. And so stay days and hospitals was converted into bigger classes

The accuracy of the used models were:



Decision Tree and KNN showed very low degree of accuracy around 50 only.

Random Forest accuracy is the highest among the 3 models used for prediction of Length of Stay.

**Chapter 5**

**Limitation and Scope of the study**

**5.1 Limitation**

One potential limitation of our study is the unavailability of a real dataset. In fact, we had to use Microsoft dataset as an example due to the difficulties that we met in access to the real data. For better results, medical experts must be involved to determine factors impacting the LOS and in the pre-processing step. The values of the LOS in the used dataset were limited to a small set. An alternative to enhance the performance of the system is to transform the LOS from a numerical data to a categorical one based on the survey, the existing work and medical experts.

The ongoing COVID-19 pandemic is itself an event which has significantly disturbed the world as a community, thus this study is also an outcome of such significant changes. This study is based on the concept of simulation modelling and it is apparent that it can be used in future as well, during other pandemics and global outbreaks to simulate the futuristic outcomes of such events on the consumers. Meanwhile, the simulation model can act as a frame of reference for other such pandemic and the communities hit with such pandemic s to determine the probable scenarios in the near future.

**5.2 Scope**

* Resources allocated to patients could be further studied which could help reduce cost of hospitals and increased efficiency.
* Given the unprecedented time of COVID 19 problem related to lack of hospital beds in India could be tackled using a similar methodology
* Similar models could be used by hotels, restaurants etc. to avoid waiting list and optimum allocation of resources.

**Bibliography**

* Awad A, Bader-El-Den M, McNicholas J. Patient length of stay and mortality prediction (2017 May;30(2)): A survey. Health Serv Manage Res.:105-120. doi: 10.1177/0951484817696212. Epub 2017 Mar 22. PMID: 28539083.
* Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S. Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. PLoS One. 2018 Apr 13;13(4):e0195901. doi: 10.1371/journal.pone.0195901. PMID: 29652932; PMCID: PMC5898738
* Baniasadi, T., Kahnouji, K., Davaridolatabadi, N., & Hosseini Teshnizi, S. (2019). Factors affecting length of stay in Children Hospital in Southern Iran. BMC health services research, 19(1), 949. <https://doi.org/10.1186/s12913-019-4799-1>
* Feibert, Diana Cordes, Jacobsen, Peter, Wallin &Michael (2017) Management Engineering Improving Healthcare Logistics Processes. Presented in DTU
* Guo A, Lu J, Tan H, Kuang Z, Luo Y, Yang T, Xu J, Yu J, Wen C, Shen A(2019) Risk factors on admission associated with hospital length of stay in patients with COVID-19: a retrospective cohort study. Sci Rep. 2021 Mar 31;11(1):7310. doi: 10.1038/s41598-021-86853-4. PMID: 33790365; PMCID: PMC8012638.
* Lagoe, R. , Drapola, B. , Luziani, M. and Pernisi, L. (2016) Hospital Length of Stay Reduction: A Long-Term Study. International Journal of Clinical Medicine, 7, 530-537. doi: 10.4236/ijcm.2016.78058.
* Rees, E.M., Nightingale, E.S., Jafari, Y. et al. COVID-19 length of hospital stay: a systematic review and data synthesis. BMC Med 18, 270 (2020). https://doi.org/10.1186/s12916-020-01726-3
* Ronald Lagoe, Louise Pernisi, Mary Luziani & Shelly Littau (2014) in Patient Length of Stay and Mortality Prediction published online showing advances in Bioscience and Biotechnology, 5, 188-196
* Stefania Tattoni, Diana Giannico, Massimiliano Schiraldi & Massimiliano Schiraldi (2019). Operations management techniques for resource optimization in health care structure in the context of radiological and diagnostic imaging services provision.
* Stone, K., Zwiggelaar, R., Jones, P., & MacParthaláin, N. (2019). Predicting Hospital Length of Stay for Accident and Emergency Admissions. 283-295. Paper presented at UK Workshop on Computational Intelligence, Portsmouth, United Kingdom of Great Britain and Northern Ireland. <https://doi.org/10.1007/978-3-030-29933-0_24>
* Tayebeh Baniasadi, Marjan Ghazisaeedi, Mehdi Hassaniazad, Sharareh R. Niakan Kalhori, Mehraban Shah (2021) . A Study of Factors Affecting on Patients’ Length of Stay in a Surgical Ward: Toward Optimization by Technology-based Solutions
* Tsai PF, Chen PC, Chen YY, Song HY, Lin HM, Lin FM, Huang QP(2016). Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. J Healthc Eng. 2016;2016:7035463. doi: 10.1155/2016/7035463. PMID: 27195660; PMCID: PMC5058566.

**References**

* <https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01726-3>
* 0<https://data.oecd.org/healthcare/length-of-hospital-stay.htm#:~:text=post%2Dacute%20settings.-,The%20ALOS%20refers%20to%20the%20average%20number%20of%20days%20that,number%20of%20admissions%20or%20discharges>.
* <https://www.worldometers.info/coronavirus/>
* https://www.worldometers.info/coronavirus/country/india/
* <https://towardsdatascience.com/label-encoder-and-onehot-encoder-in-python-83d32288b592>
* <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>
* <http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
* <http://scikit-learn.org/stable/modules/tree.html>
* <http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
* https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6905070/>
* <https://pure.aber.ac.uk/portal/en/publications/predicting-hospital-length-of-stay-for-accident-and-emergency-admissions(c8e2b9d4-2878-4ca6-935e-c163563f37d4).html>
* <https://www.researchgate.net/publication/340193564_A_Study_of_Factors_Affecting_on_Patients'_Length_of_Stay_in_a_Surgical_Ward_Toward_Optimization_by_Technology-based_Solutions>
* <https://www.scirp.org/journal/paperabs.aspx?paperid=69603>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5898738/>
* <https://pubmed.ncbi.nlm.nih.gov/33790365/>
* <https://www.researchgate.net/publication/251734551_Operations_management_techniques_for_resource_optimization_in_health_care_structure>
* <https://www.nature.com/articles/s41598-021-86853-4>
* <https://www.scirp.org/journal/paperabs.aspx?paperid=69603>
* https://www.researchgate.net/publication/251734551\_Operations\_management\_techniques\_for\_resource\_optimization\_in\_health\_care\_structure
* <https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01726-3>
* <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207203>
* <https://www.sciencedirect.com/science/article/pii/S2212567116303203>
* <https://www.oecd-ilibrary.org/sites/0d8bb30a-en/index.html?itemId=/content/component/0d8bb30a-en>

**Annexure A**

* **Applications Used**
* <https://research.google.com/colaboratory/>
* <https://www.kaggle.com/>
* Anaconda: Jupiter
* Microsoft Word
* **Libraries Imported**

