Data Science Project

Bachelors of Engineering in Computer Engineering

Submitted By

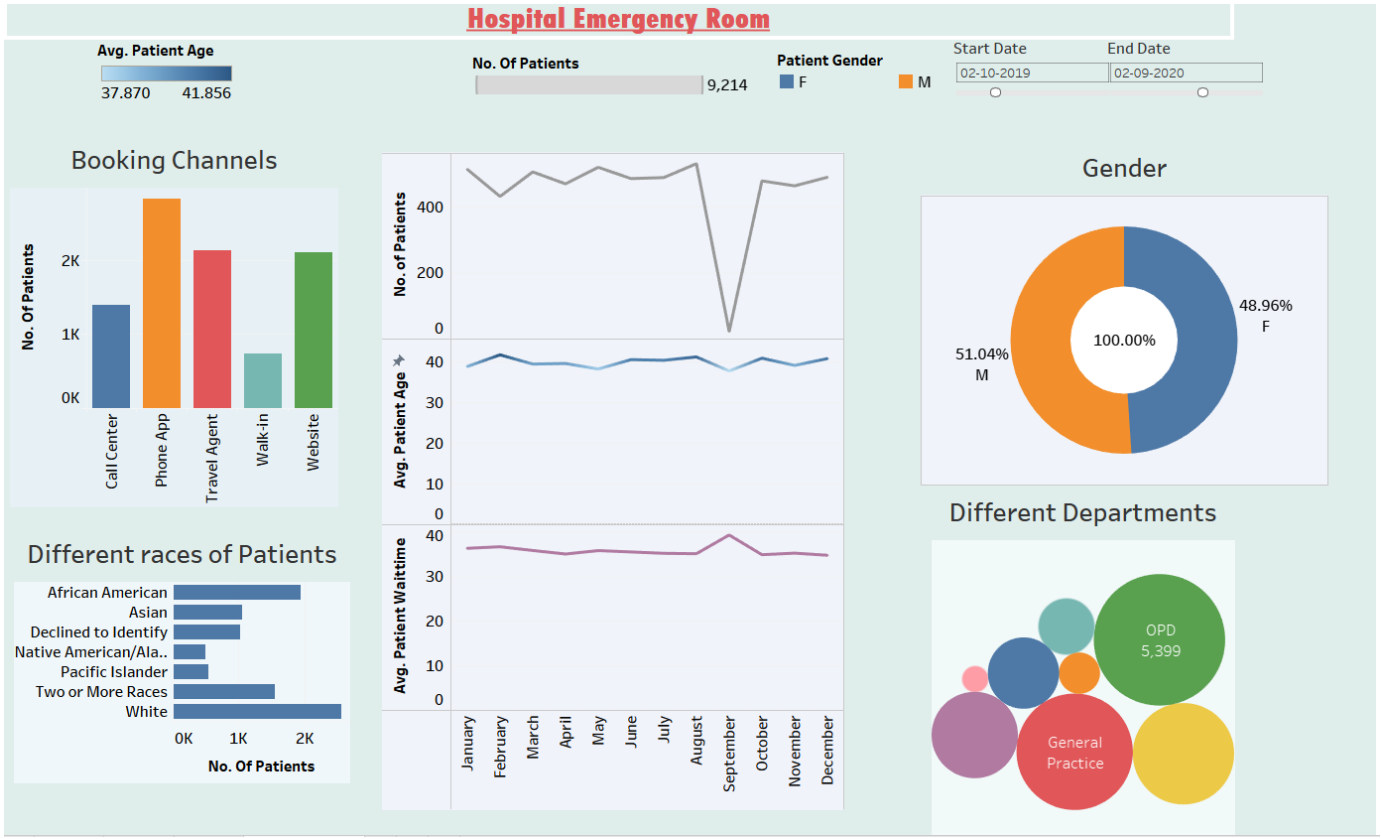
Avneet Kaur 102003487

Department of Computer Science Engineering

Submitted to: Mrs. Kashish Goyal

Thapar Institute of Engineering and Technology, Patiala

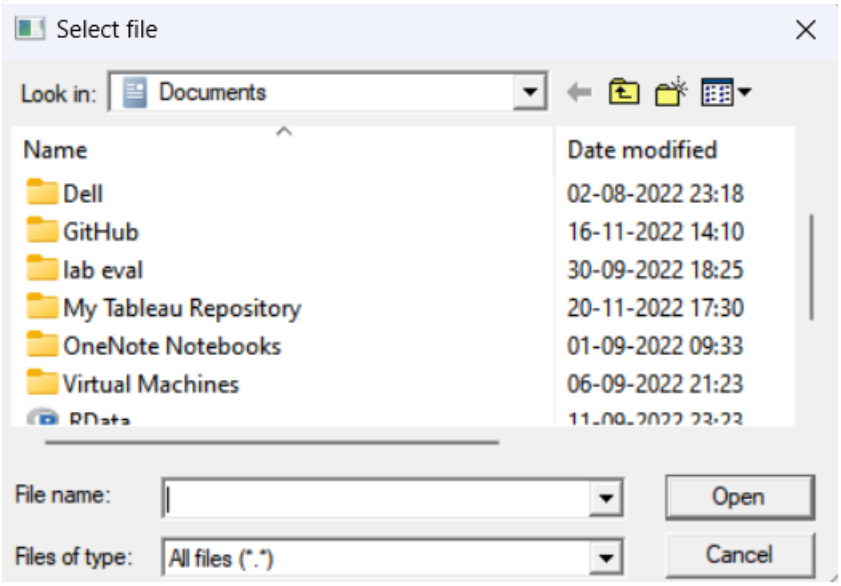
Dashboard:



Data Pre-processing:

1.Merging the two datasets

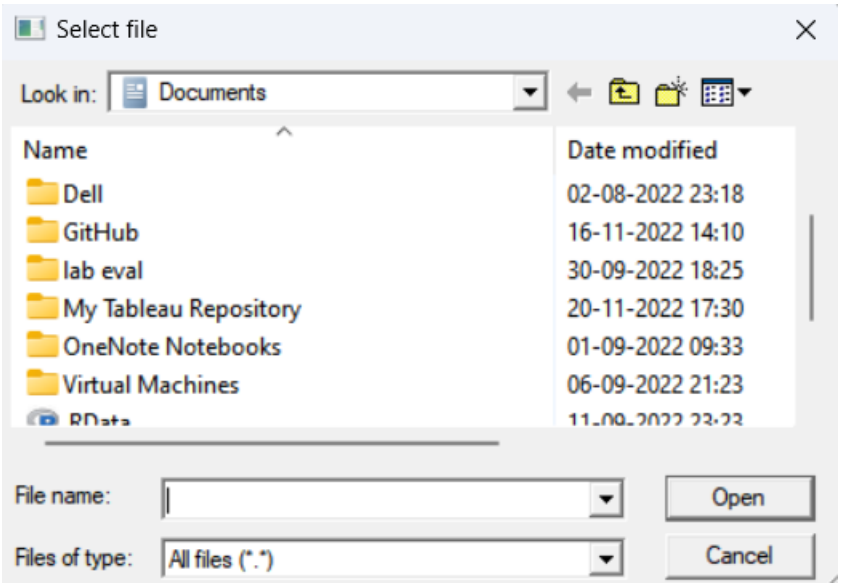
```
library(dplyr)
#Choose table1 from pc

table1<-read.csv(file.choose())
```



```
#add the column which is having weekday extracted from date time column
z <- weekdays(as.Date(table1$date, "%Y-%m-%d"))
table1$Day<-z
z
```

```
  [1] "Friday"    "Monday"    "Saturday"  "Tuesday"   "Friday"    "Saturday"  "Friday"    "Monday"    "Wednesday" "Sunday"    "Sunday"
 [12] "Tuesday"   "Tuesday"   "Tuesday"   "Saturday"  "Thursday"  "Monday"    "Tuesday"   "Wednesday" "Saturday"  "Sunday"    "Wednesday"
 [23] "Sunday"    "Sunday"    "Friday"    "Friday"    "Friday"    "Monday"    "Wednesday" "Tuesday"   "Thursday"  "Sunday"    "Wednesday"
 [34] "Wednesday" "Thursday"  "Wednesday" "Saturday"  "Sunday"    "Monday"    "Saturday"  "Monday"    "Saturday"  "Monday"    "Monday"
 [45] "Monday"    "Tuesday"   "Monday"    "Monday"    "Friday"    "Thursday"  "Tuesday"   "Tuesday"   "Friday"    "Friday"    "Saturday"
 [56] "Thursday"  "Wednesday" "Thursday"  "Tuesday"   "Thursday"  "Monday"    "Tuesday"   "Sunday"    "Thursday"  "Friday"    "Wednesday"
 [67] "Friday"    "Thursday"  "Monday"    "Saturday"  "Sunday"    "Thursday"  "Monday"    "Sunday"    "Thursday"  "Wednesday" "Friday"
 [78] "Monday"    "Saturday"  "Wednesday" "Friday"    "Monday"    "Sunday"    "Saturday"  "Sunday"    "Wednesday" "Friday"    "Thursday"
 [89] "Tuesday"   "Monday"    "Tuesday"   "Friday"    "Sunday"    "Saturday"  "Wednesday" "Friday"    "Monday"    "Friday"    "Monday"
100] "Wednesday" "Friday"    "Sunday"    "Tuesday"   "Monday"    "Tuesday"   "Thursday"  "Wednesday" "Tuesday"   "Monday"    "Sunday"
```

```
# choose table 2 from pc
table2<-read.csv(file.choose())
```



2.Selecting number of rows in table 2 equal to table 1 as we need to merge two of them so row count should be same.

```
#choose rows
table2<-head(table2,9216)
class(table1)
```

```
[1] "data.frame"
```

3.Creating a dataframe named table 3

```
table3<-data.frame("patient_id" = table1$patient_id ,"stay_duration" = table2$stay_duration , "booking_channel"= table2$booking_channel)
final_table<-merge.data.frame(table1,table3,by='patient_id')
```

```
    patient_id                date patient_gender patient_age patient_sat_score patient_first_inital patient_last_name
1  100-04-3993 2019-04-04 04:50:19              F          29                NA                    M            St Ange
2  100-17-5081 2020-01-14 19:20:06              M          67                NA                    V            Flicker
3  100-21-9648 2020-01-17 18:53:09              F          39                 8                    W             Marran
4  100-34-6753 2020-05-13 14:03:28              M          43                NA                    B             Paulus
5  100-34-9587 2020-04-01 04:17:42              M          20                NA                    U            Lamburn
6  100-40-2709 2020-05-08 04:58:40              M          77                NA                    O            Cammack
7  100-66-0896 2020-03-26 12:14:29              M           2                 6                    I           Prickett
8  100-66-8222 2019-12-23 12:07:40              F          65                NA                    F            Mullane
9  100-67-1276 2019-11-03 08:26:14              M          55                10                    S           Hallbird
10 100-70-0071 2020-01-14 01:43:19              M          38                NA                    R            Downham
11 100-72-5705 2020-06-19 22:16:51              F          60                NA                    N              Dudny
12 100-74-3943 2019-09-17 01:31:02              F           3                 3                    M            Hallard

                   patient_race patient_admin_flag patient_waittime department_referral       Day stay_duration booking_channel
1                         White              false              16                None  Thursday             7         Website
2              African American              false              60                None   Tuesday             7       Phone App
3              Pacific Islander               true              22                None    Friday             9       Phone App
4              Pacific Islander               true              25    General Practice Wednesday            12    Call Center
5           Declined to Identify              false              24          Neurology Wednesday             2         Walk-in
6                         White              false              48                None    Friday             1         Website
7              African American               true              23         Orthopedics  Thursday            11    Call Center
8                         Asian              false              17    General Practice    Monday            11       Phone App
9                         White               true              11         Orthopedics    Sunday            12       Phone App
10             African American              false              57                None   Tuesday             6         Website
11             African American               true              45                None    Friday             7    Call Center
12                        White               true              14                None   Tuesday            12       Phone App
```

4.Finding missing values:

```
#count of missing values
sum(is.na(final_table))
sum(is.na(final_table))
L] 6699
```

5.Removing missing values:

```
#dropping the column with null values in final table
final_table<-final_table[-5]
```

6..Reading the new file having deleted column

```
#writing the file
write.csv(final_table,"C:/Users/Avneet kaur/Downloads/final_table.csv")
```

7.Editting the table

```
#editting the table
final_table=edit(final_table)
```

	patient_id	date	patient_gender	patient_age
1	100-04-3993	2019-04-04 04:50:19	F	29
2	100-17-5081	2020-01-14 19:20:06	M	67
3	100-21-9648	2020-01-17 18:53:09	F	39
4	100-34-6753	2020-05-13 14:03:28	M	43
5	100-34-9587	2020-04-01 04:17:42	M	20
6	100-40-2709	2020-05-08 04:58:40	M	77
7	100-66-0896	2020-03-26 12:14:29	M	2
8	100-66-8222	2019-12-23 12:07:40	F	65
9	100-67-1276	2019-11-03 08:26:14	M	55
10	100-70-0071	2020-01-14 01:43:19	M	38
11	100-72-5705	2020-06-19 22:16:51	F	60
12	100-74-3943	2019-09-17 01:31:02	F	3
13	100-74-5636	2020-08-13 06:07:22	F	47
14	100-79-0109	2020-02-27 04:51:45	F	19
15	100-81-9769	2020-03-09 19:26:17	M	28
16	100-84-7203	2019-06-13 21:33:09	F	37
17	101-08-8798	2020-07-31 14:18:46	F	72
18	101-13-4808	2019-04-25 01:36:55	F	30
19	101-35-3930	2020-01-24 07:29:39	F	30

8.Removing the negative values

```
#Removing the negative values you inserted through editor
final_table$patient_waittime<- abs(final_table$patient_waittime)
final_table
```

	patient_admin_flag	patient_waittime	department_referral	Day	stay_duration	booking_channel
1	false	90	None	Thursday	7	Website
2	false	60	None	Tuesday	7	Phone App
3	true	22	None	Friday	9	Phone App
4	true	25	General Practice	Wednesday	12	Call Center
5	false	24	Neurology	Wednesday	2	Walk-in
6	false	48	None	Friday	1	Website
7	true	23	Orthopedics	Thursday	11	Call Center
8	false	17	General Practice	Monday	11	Phone App
9	true	11	Orthopedics	Sunday	12	Phone App
10	false	57	None	Tuesday	6	Website
11	true	45	None	Friday	7	Call Center
12	true	14	None	Tuesday	12	Phone App

Dataset Information:

```
#dimensions
dim(final_table)
```

```
[1] 9216   13
```

Dataset Analysis:

1. Finding distinct data:

```
#count of male and female
nrow(subset(final_table,patient_gender=="M"))
nrow(subset(final_table,patient_gender=="F"))
```

```
 nrow(subset(final_table,patient_gender=="M"))
1] 4705
 nrow(subset(final_table,patient_gender=="F"))
1] 4487
```

2. Splitting on department basis

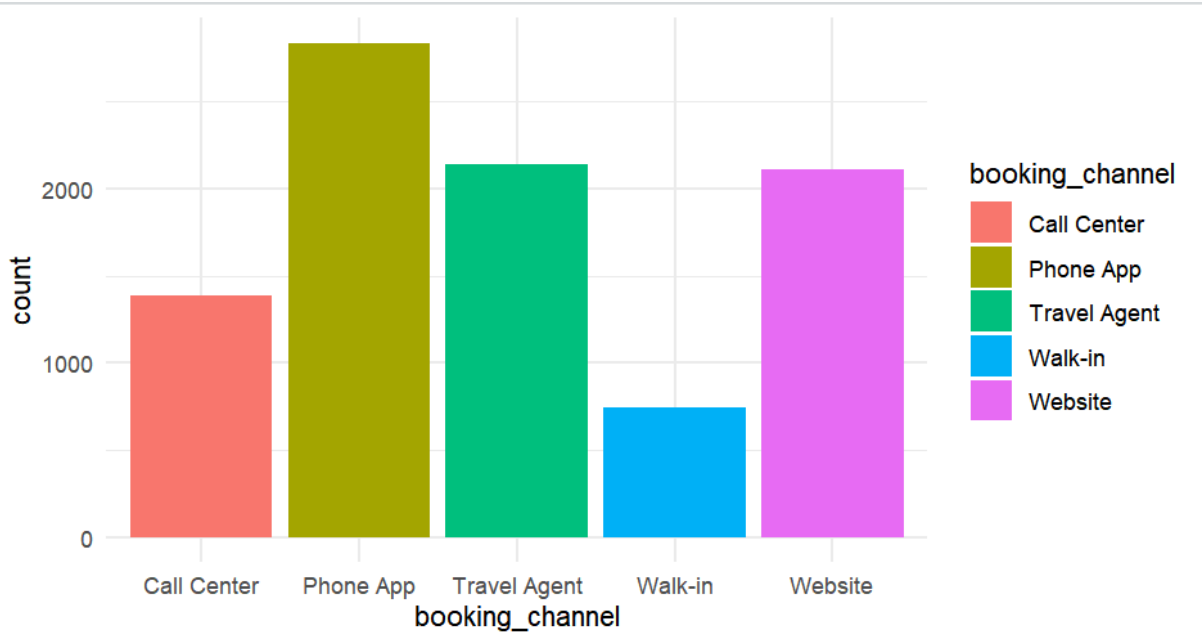
```
#splitting final_table on basis of department
split(final_table,final_table$department_referral)
```

	patient_admin_flag	patient_waittime	department_referral	Day	stay_duration	booking_channel
113	true	59	Renal	Tuesday	11	Website
390	true	47	Renal	Wednesday	3	Walk-in
396	false	42	Renal	Wednesday	3	Website
447	true	55	Renal	Monday	10	Travel Agent
486	true	49	Renal	Sunday	1	Walk-in
588	true	43	Renal	Tuesday	13	Call Center
677	false	12	Renal	Friday	2	Phone App
1065	false	52	Renal	Friday	8	Call Center
1147	false	20	Renal	Tuesday	11	Walk-in
1152	true	42	Renal	Thursday	10	Website
1264	true	31	Renal	Tuesday	14	Walk-in
1312	true	47	Renal	Tuesday	7	Phone App

3. Plotting between booking channel and count

```
#plot
install.packages("ggplot2")
library(ggplot2)
ggplot(final_table, aes(x =booking_channel,fill=booking_channel)) +geom_bar()+theme_minimal()
```

```
1741              true                27    Gastroenterology Wednesday
        1     Call Center
1743              true                18    Gastroenterology    Sunday
        8       Phone App
1792             false                44    Gastroenterology Wednesday
        4       Website
1843             false                54    Gastroenterology  Thursday
        4       Phone App
1855             false                57    Gastroenterology    Friday
        5       Phone App
1881              true                14    Gastroenterology   Tuesday
        2       Phone App
1898              true                58    Gastroenterology    Sunday
```



4. Unique values in column

```
#unique categories in column
print(unique(final_table$patient_race))
```

```
[1] "White"                          "African American"
[3] "Pacific Islander"               "Declined to Identify"
[5] "Asian"                          "Two or More Races"
[7] "Native American/Alaska Native"
```

5. Creating table

```
#table
table(final_table$department_referral)
```

```
  Cardiology Gastroenterology General Practice         Neurology
         248              178              1840               193
        None      Orthopedics     Physiotherapy             Renal
        5400              995               276                86
```