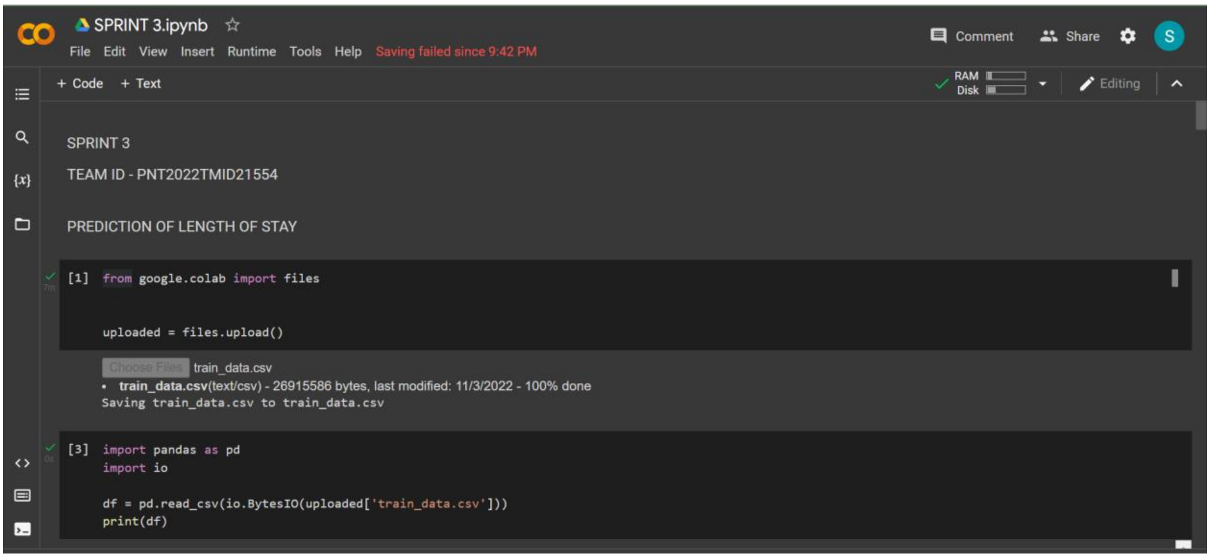
Sprint 3

TEAM ID - PNT2022TMID21554

Prediction of Length of Stay

1.Uploading Necessary files


SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help Saving failed since 9:42 PM
Comment Share
+ Code + Text
RAM Disk
Editing
Choose Files

SPRINT 3

TEAM ID - PNT2022TMID21554

PREDICTION OF LENGTH OF STAY

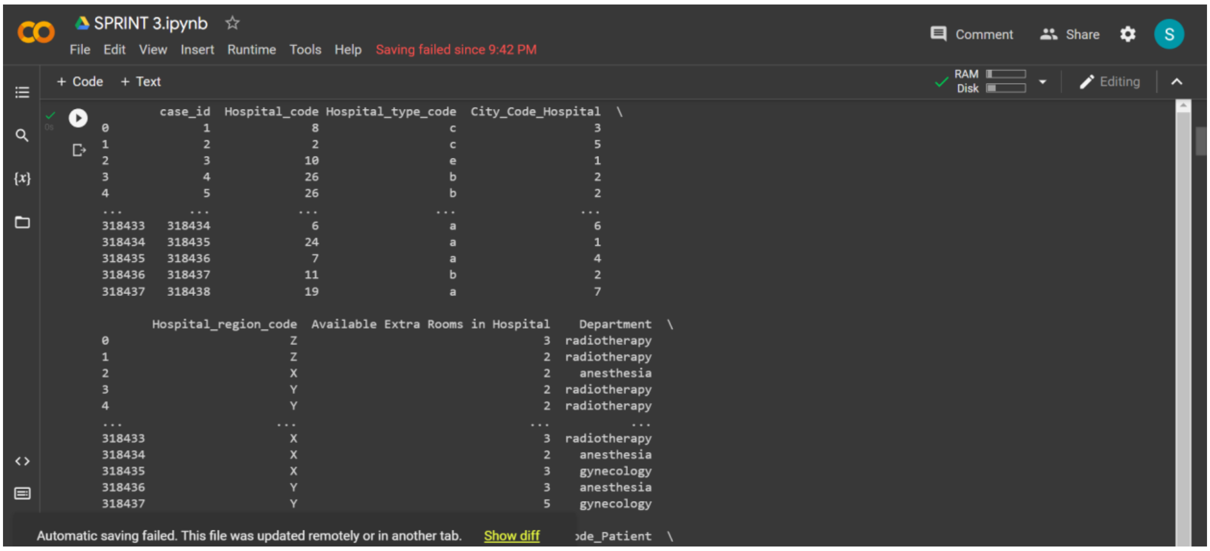
```
[1] from google.colab import files


    uploaded = files.upload()
```

train_data.csv

- **train_data.csv**(text/csv) - 26915586 bytes, last modified: 11/3/2022 - 100% done

```
Saving train_data.csv to train_data.csv
```

```
[3] import pandas as pd
    import io

    df = pd.read_csv(io.BytesIO(uploaded['train_data.csv']))
    print(df)
```

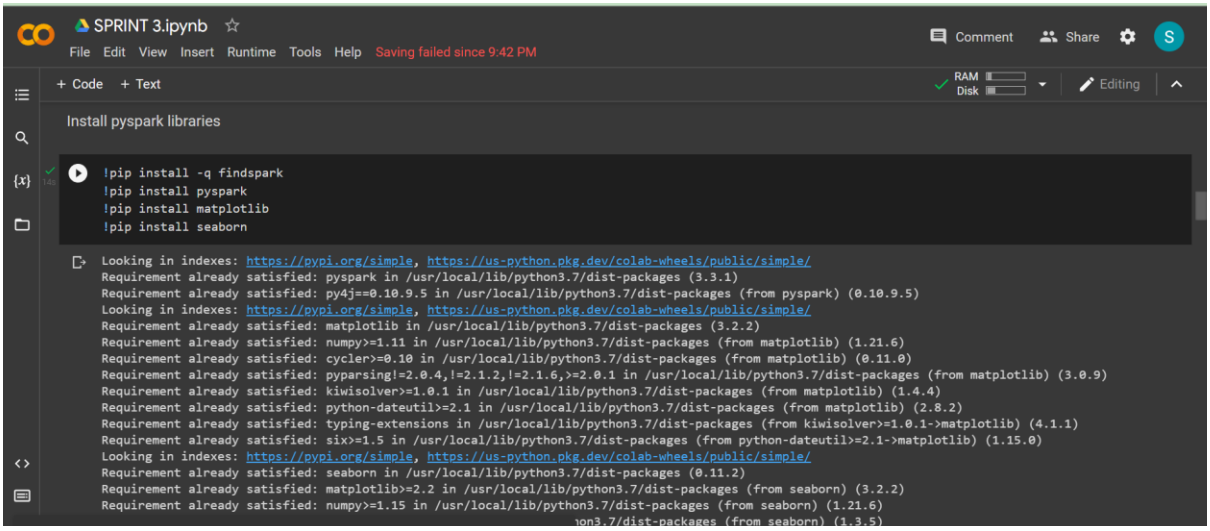

SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help Saving failed since 9:42 PM
Comment Share
+ Code + Text
RAM Disk
Editing

```
        case_id  Hospital_code Hospital_type_code  City_Code_Hospital  \
0             1              8                  c                   3
1             2              2                  c                   5
2             3             10                  e                   1
3             4             26                  b                   2
4             5             26                  b                   2
...         ...            ...                ...                 ...
318433   318434              6                  a                   6
318434   318435             24                  a                   1
318435   318436              7                  a                   4
318436   318437             11                  b                   2
318437   318438             19                  a                   7

       Hospital_region_code  Available Extra Rooms in Hospital    Department  \
0                         Z                                  3  radiotherapy
1                         Z                                  2  radiotherapy
2                         X                                  2    anesthesia
3                         Y                                  2  radiotherapy
4                         Y                                  2  radiotherapy
...                     ...                                ...           ...
318433                    X                                  3  radiotherapy
318434                    X                                  2    anesthesia
318435                    X                                  3    gynecology
318436                    Y                                  3    anesthesia
318437                    Y                                  5    gynecology
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
ode_Patient  \
```

Installing pyspark Libraries

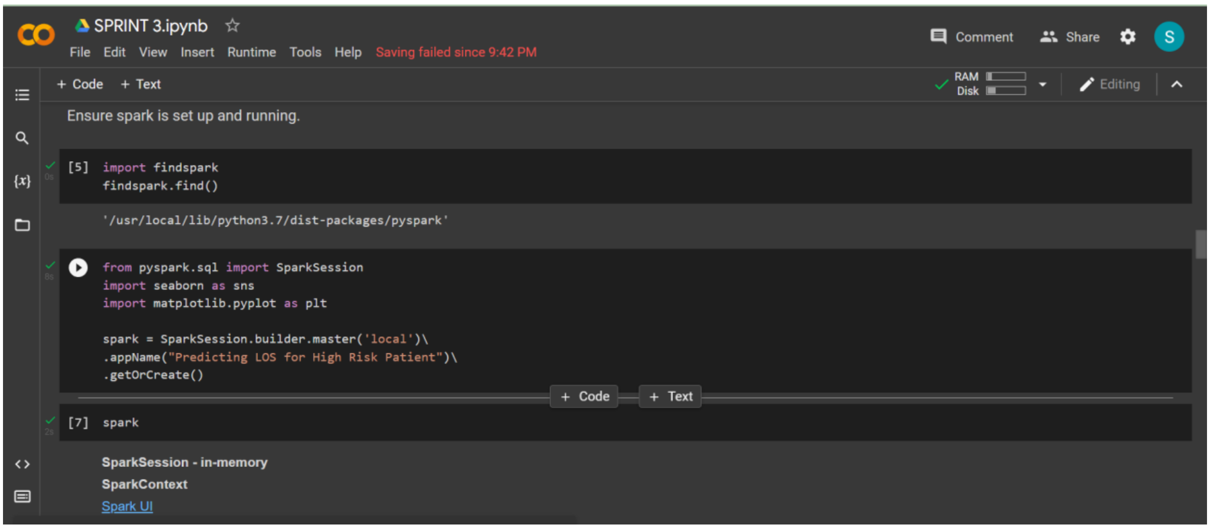

SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help Saving failed since 9:42 PM
Comment
Share
S
+ Code + Text
RAM
Disk
Editing

Install pyspark libraries

```
!pip install -q findspark
!pip install pyspark
!pip install matplotlib
!pip install seaborn
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pyspark in /usr/local/lib/python3.7/dist-packages (3.3.1)
Requirement already satisfied: py4j==0.10.9.5 in /usr/local/lib/python3.7/dist-packages (from pyspark) (0.10.9.5)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (3.2.2)
Requirement already satisfied: numpy>=1.11 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (1.21.6)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (0.11.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (3.0.9)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (1.4.4)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from kiwisolver>=1.0.1->matplotlib) (4.1.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.1->matplotlib) (1.15.0)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: seaborn in /usr/local/lib/python3.7/dist-packages (0.11.2)
Requirement already satisfied: matplotlib>=2.2 in /usr/local/lib/python3.7/dist-packages (from seaborn) (3.2.2)
Requirement already satisfied: numpy>=1.15 in /usr/local/lib/python3.7/dist-packages (from seaborn) (1.21.6)
                                                                  on3.7/dist-packages (from seaborn) (1.3.5)
```

Ensuring spark is setup and it is running


SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help Saving failed since 9:42 PM
Comment
Share
S
+ Code + Text
RAM
Disk
Editing
+ Code
+ Text

Ensure spark is set up and running.

```
[5] import findspark
    findspark.find()
```

```
'/usr/local/lib/python3.7/dist-packages/pyspark'
```

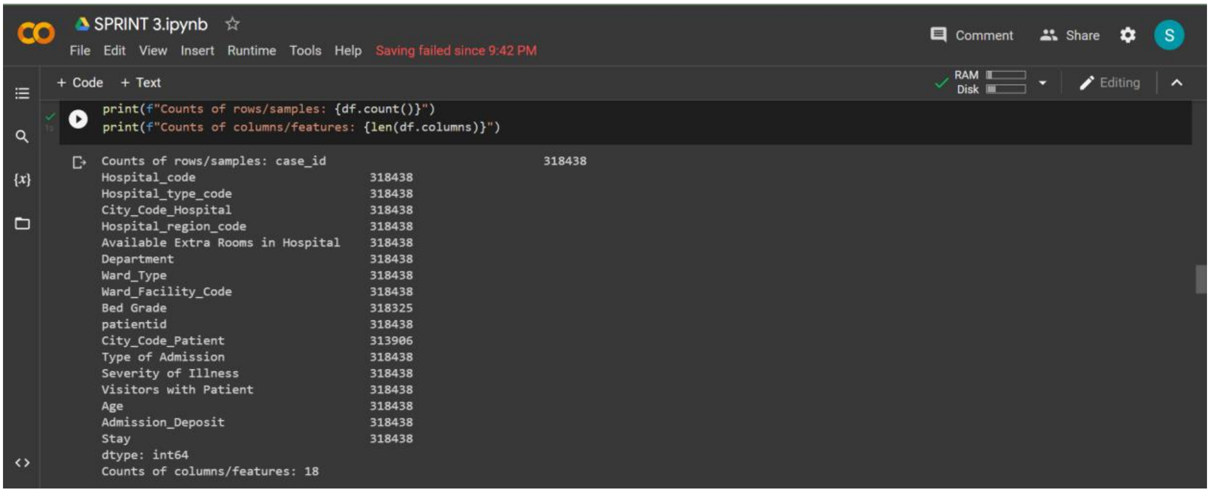
```
from pyspark.sql import SparkSession
import seaborn as sns
import matplotlib.pyplot as plt

spark = SparkSession.builder.master('local')\
.appName("Predicting LOS for High Risk Patient")\
.getOrCreate()
```

```
[7] spark
```

SparkSession - in-memory

SparkContext

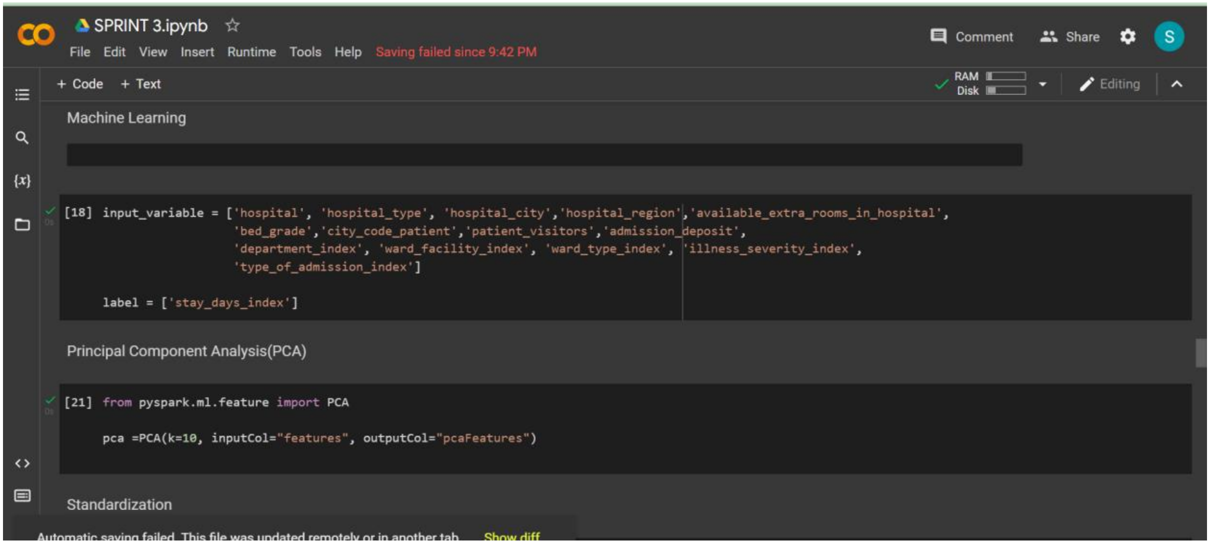
Spark UI


SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help Saving failed since 9:42 PM
Comment
Share
S
+ Code + Text
RAM
Disk
Editing

```
print(f"Counts of rows/samples: {df.count()}")
print(f"Counts of columns/features: {len(df.columns)}")
```

```
Counts of rows/samples: case_id                             318438
Hospital_code                     318438
Hospital_type_code                318438
City_Code_Hospital                318438
Hospital_region_code              318438
Available Extra Rooms in Hospital 318438
Department                        318438
Ward_Type                         318438
Ward_Facility_Code                318438
Bed Grade                         318325
patientid                         318438
City_Code_Patient                 313906
Type of Admission                 318438
Severity of Illness               318438
Visitors with Patient             318438
Age                               318438
Admission_Deposit                 318438
Stay                              318438
dtype: int64
Counts of columns/features: 18
```

Using Machine Learning Algorithm and Principal Component Analysis(PCA)


SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help Saving failed since 9:42 PM
Comment Share
+ Code + Text
RAM Disk
Editing

Machine Learning

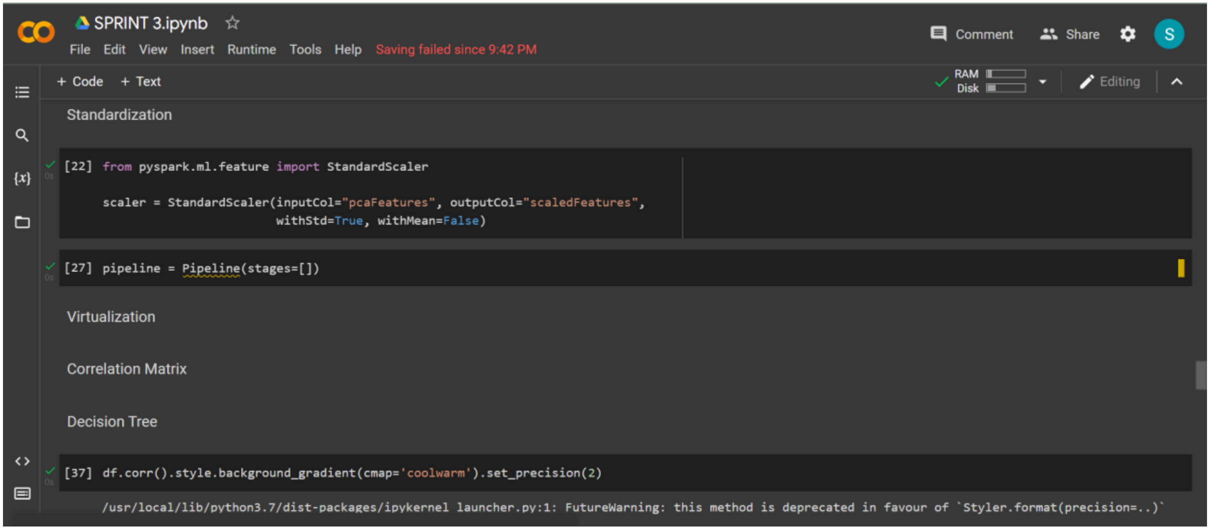
```
[18] input_variable = ['hospital', 'hospital_type', 'hospital_city','hospital_region','available_extra_rooms_in_hospital',
                      'bed_grade','city_code_patient','patient_visitors','admission_deposit',
                      'department_index', 'ward_facility_index', 'ward_type_index', 'illness_severity_index',
                      'type_of_admission_index']

     label = ['stay_days_index']
```

Principal Component Analysis(PCA)

```
[21] from pyspark.ml.feature import PCA

     pca =PCA(k=10, inputCol="features", outputCol="pcaFeatures")
```

Standardization

Automatic saving failed. This file was updated remotely or in another tab. Show diff


SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help Saving failed since 9:42 PM
Comment Share
+ Code + Text
RAM Disk
Editing

Standardization

```
[22] from pyspark.ml.feature import StandardScaler

     scaler = StandardScaler(inputCol="pcaFeatures", outputCol="scaledFeatures",
                             withStd=True, withMean=False)
```

```
[27] pipeline = Pipeline(stages=[])
```

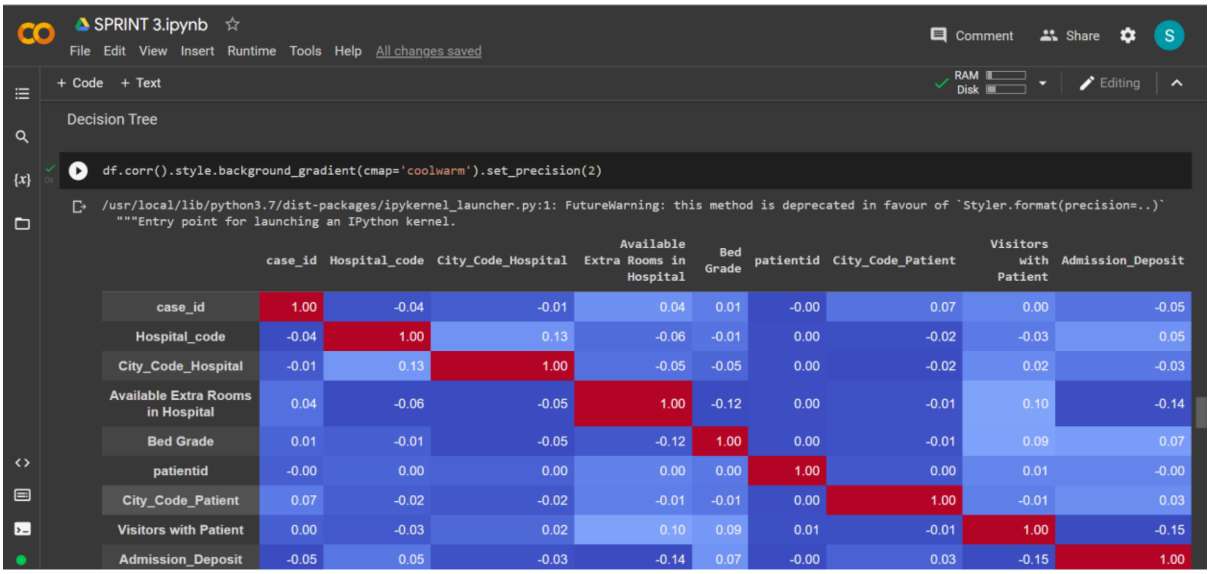
Virtualization

Correlation Matrix

Decision Tree

```
[37] df.corr().style.background_gradient(cmap='coolwarm').set_precision(2)

     /usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: this method is deprecated in favour of `Styler.format(precision=..)`
```

Decision Tree Algorithm for prediction

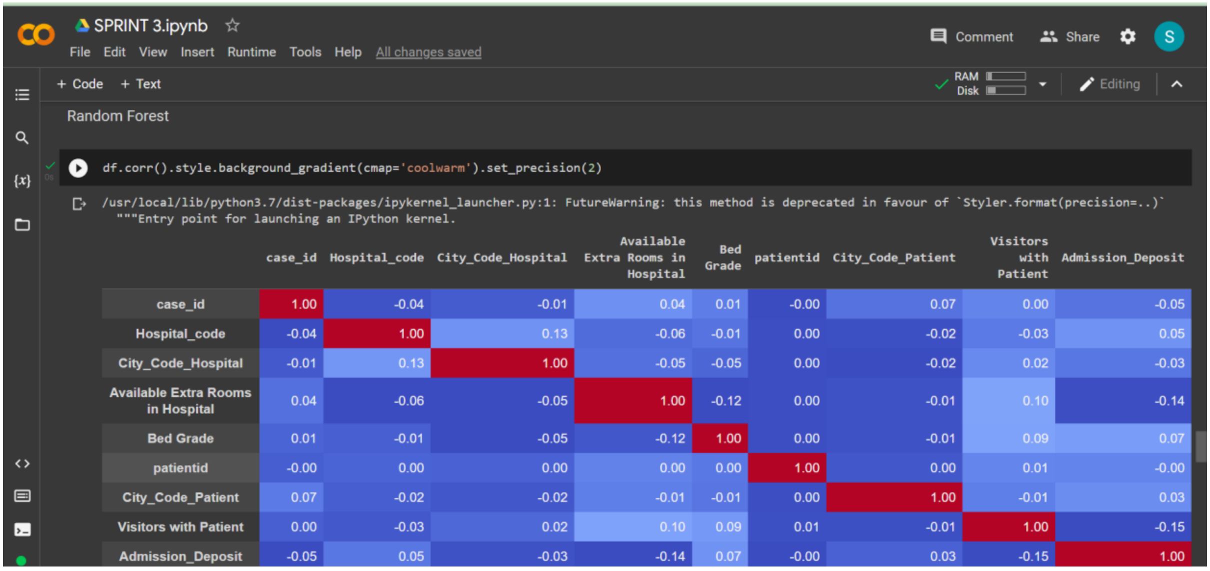


Decision Tree

```
df.corr().style.background_gradient(cmap='coolwarm').set_precision(2)

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: this method is deprecated in favour of `Styler.format(precision=..)`
  """Entry point for launching an IPython kernel.
```

	case_id	Hospital_code	City_Code_Hospital	Available Extra Rooms in Hospital	Bed Grade	patientid	City_Code_Patient	Visitors with Patient	Admission_Deposit
case_id	1.00	-0.04	-0.01	0.04	0.01	-0.00	0.07	0.00	-0.05
Hospital_code	-0.04	1.00	0.13	-0.06	-0.01	0.00	-0.02	-0.03	0.05
City_Code_Hospital	-0.01	0.13	1.00	-0.05	-0.05	0.00	-0.02	0.02	-0.03
Available Extra Rooms in Hospital	0.04	-0.06	-0.05	1.00	-0.12	0.00	-0.01	0.10	-0.14
Bed Grade	0.01	-0.01	-0.05	-0.12	1.00	0.00	-0.01	0.09	0.07
patientid	-0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.01	-0.00
City_Code_Patient	0.07	-0.02	-0.02	-0.01	-0.01	0.00	1.00	-0.01	0.03
Visitors with Patient	0.00	-0.03	0.02	0.10	0.09	0.01	-0.01	1.00	-0.15
Admission_Deposit	-0.05	0.05	-0.03	-0.14	0.07	-0.00	0.03	-0.15	1.00

Random Forest



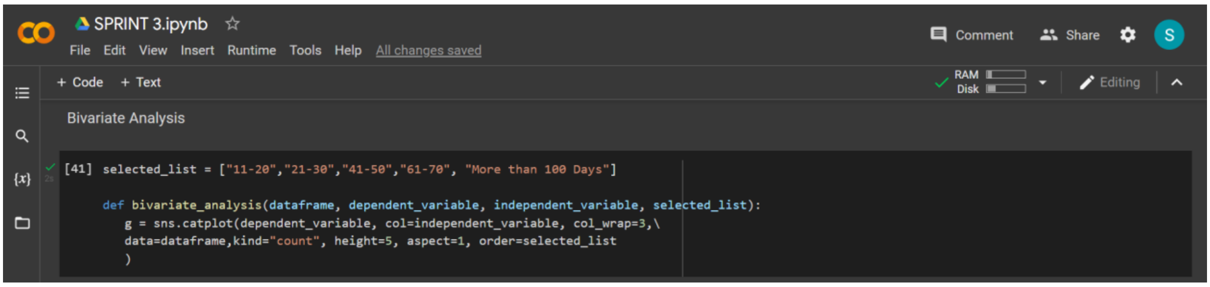
Random Forest

```
df.corr().style.background_gradient(cmap='coolwarm').set_precision(2)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: this method is deprecated in favour of `Styler.format(precision=..)`
  """Entry point for launching an IPython kernel.
```

	case_id	Hospital_code	City_Code_Hospital	Available Extra Rooms in Hospital	Bed Grade	patientid	City_Code_Patient	Visitors with Patient	Admission_Deposit
case_id	1.00	-0.04	-0.01	0.04	0.01	-0.00	0.07	0.00	-0.05
Hospital_code	-0.04	1.00	0.13	-0.06	-0.01	0.00	-0.02	-0.03	0.05
City_Code_Hospital	-0.01	0.13	1.00	-0.05	-0.05	0.00	-0.02	0.02	-0.03
Available Extra Rooms in Hospital	0.04	-0.06	-0.05	1.00	-0.12	0.00	-0.01	0.10	-0.14
Bed Grade	0.01	-0.01	-0.05	-0.12	1.00	0.00	-0.01	0.09	0.07
patientid	-0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.01	-0.00
City_Code_Patient	0.07	-0.02	-0.02	-0.01	-0.01	0.00	1.00	-0.01	0.03
Visitors with Patient	0.00	-0.03	0.02	0.10	0.09	0.01	-0.01	1.00	-0.15
Admission_Deposit	-0.05	0.05	-0.03	-0.14	0.07	-0.00	0.03	-0.15	1.00

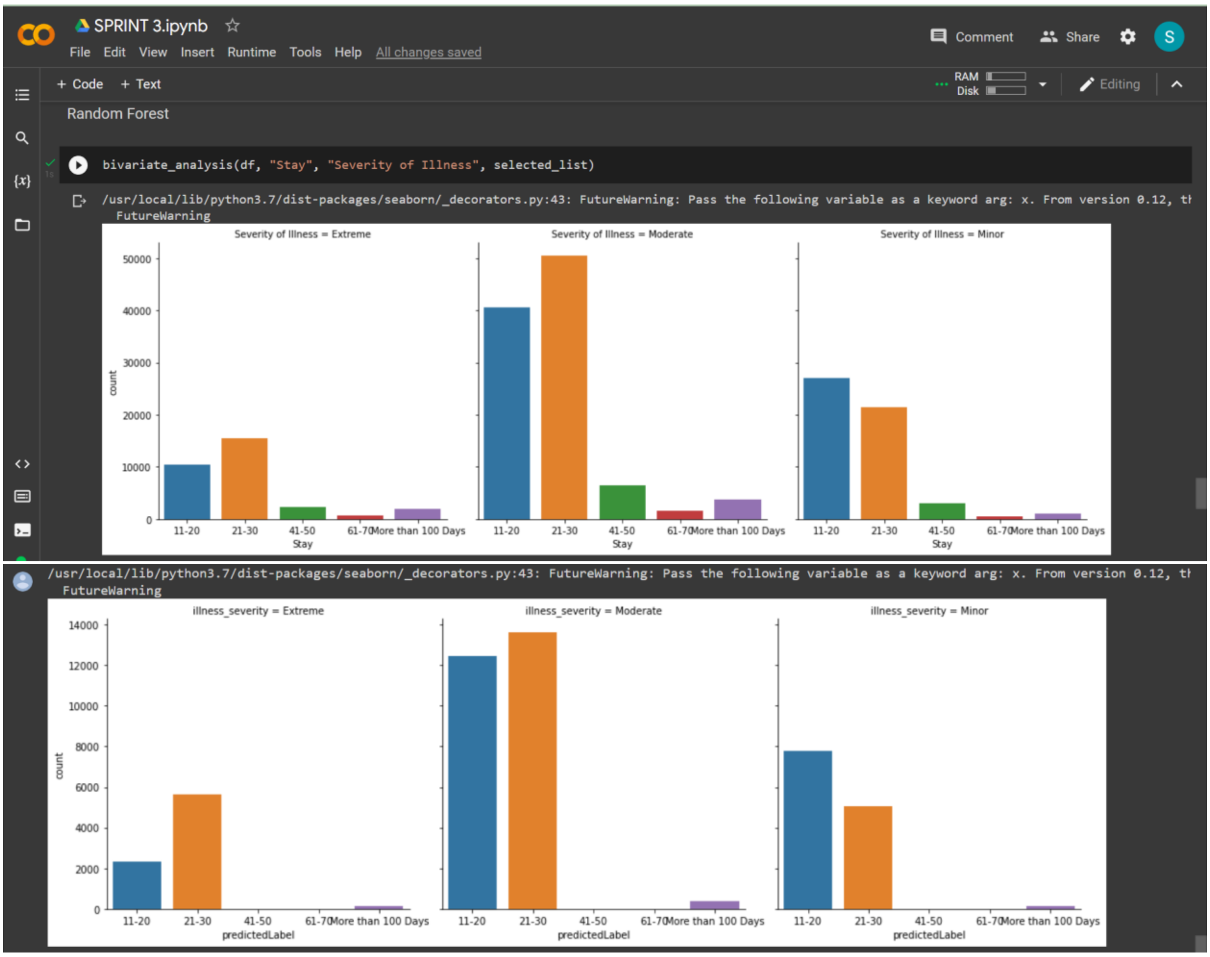
Bivariate Analysis



Bivariate Analysis

```
[41] selected_list = ["11-20","21-30","41-50","61-70", "More than 100 Days"]

     def bivariate_analysis(dataframe, dependent_variable, independent_variable, selected_list):
        g = sns.catplot(dependent_variable, col=independent_variable, col_wrap=3,\
        data=dataframe,kind="count", height=5, aspect=1, order=selected_list
        )
```

Random Forest to predict Length of Stay



SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share S
+ Code + Text
RAM Disk
Editing
Random Forest

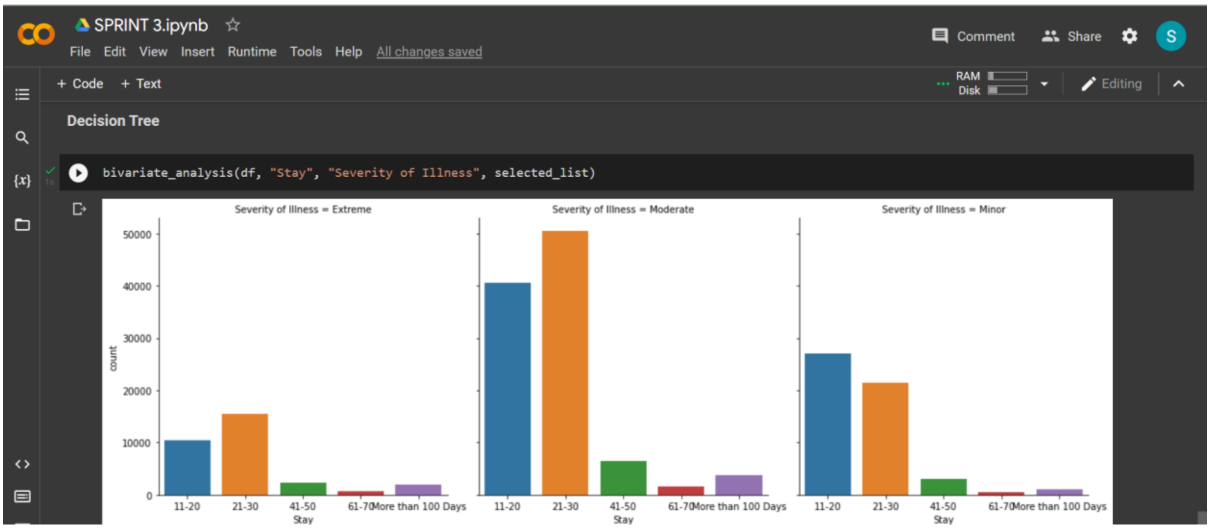
```
bivariate_analysis(df, "Stay", "Severity of Illness", selected_list)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, tl
  FutureWarning
```

Severity of Illness = Extreme
Severity of Illness = Moderate
Severity of Illness = Minor
count
0
10000
20000
30000
40000
50000
11-20
21-30
41-50
61-70More than 100 Days
Stay

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, tl
  FutureWarning
```

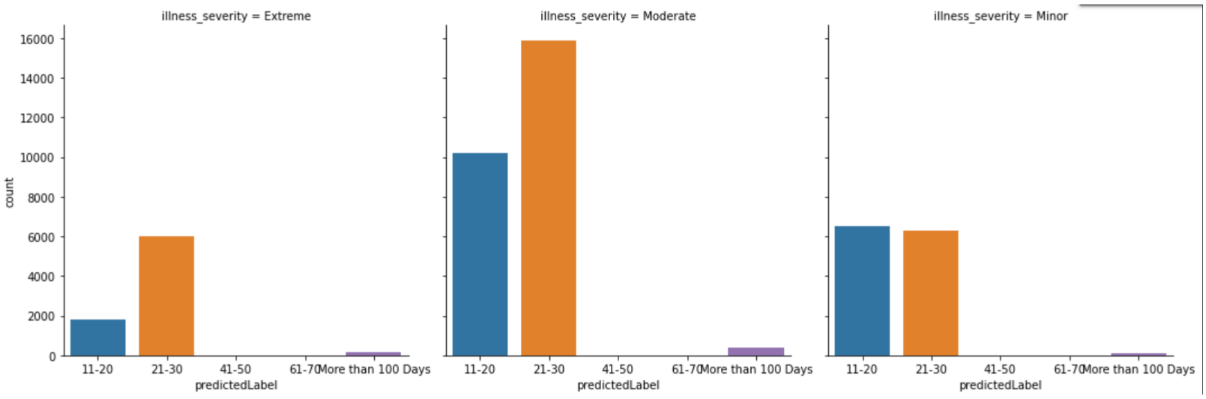
Decision Tree to predict Length of Stay



SPRINT 3.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share S
+ Code + Text
RAM Disk
Editing
Decision Tree

```
bivariate_analysis(df, "Stay", "Severity of Illness", selected_list)
```

Severity of Illness = Extreme
Severity of Illness = Moderate
Severity of Illness = Minor
count
0
10000
20000
30000
40000
50000
11-20
21-30
41-50
61-70More than 100 Days
Stay



Link

https://colab.research.google.com/drive/1Z6ZB9STV8FzW3ry-uNE0jDvupfpJxPLq#scrollTo=UPlncaqRxU8d