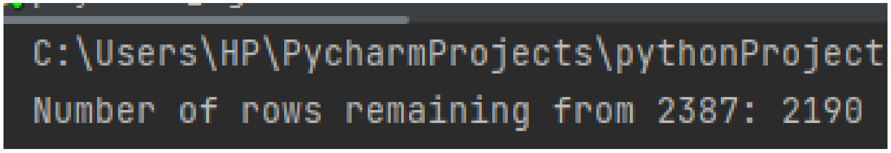
# Milestone 1 Machine Learning
# Doctor’s Fee Prediction

| | | |
|---|---|---|
| Mohamed Yasser Abd El-fattah | 2021170943 | Rob3 |
| Al Hussien Adel | 2021170941 | Rob3 |
| Mohamed Hossam Mohamed | 20201701751 | AI3 |
| Mena Mored ibrahem | 20161701065 | Bio4 |
| Mahmoud Tarek Mahfouz | 20201701752 | AI3 |

# Pre-Processing:

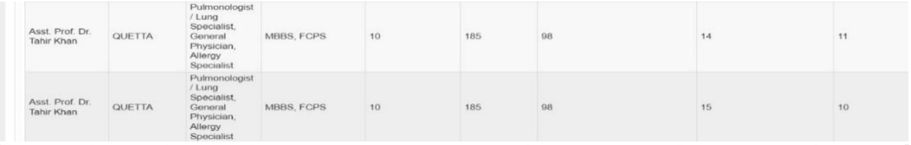
## 1. Data Cleaning

- **The dataset was deduplicated based on the "Doctor Name" column by keeping only the first occurrence of each doctor's name and removing the rest. This approach ensured each doctor's name appeared only once in the dataset for a cleaner, more accurate representation**

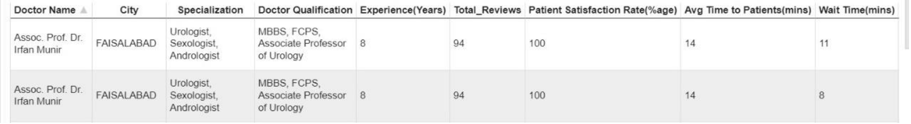


```
C:\Users\HP\PycharmProjects\pythonProject
Number of rows remaining from 2387: 2190
```

**Example for Duplicates:**

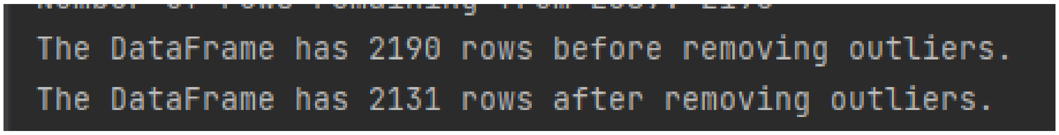


| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Asst. Prof. Dr. Tahir Khan | QUETTA | Pulmonologist / Lung Specialist, General Physician, Allergy Specialist | MBBS, FCPS | 10 | 185 | 98 | 14 | 11 |
| Asst. Prof. Dr. Tahir Khan | QUETTA | Pulmonologist / Lung Specialist, General Physician, Allergy Specialist | MBBS, FCPS | 10 | 185 | 98 | 15 | 10 |



| Doctor Name ▲ | City | Specialization | Doctor Qualification | Experience(Years) | Total_Reviews | Patient Satisfaction Rate(%age) | Avg Time to Patients(mins) | Wait Time(mins) |
|---|---|---|---|---|---|---|---|---|
| Assoc. Prof. Dr. Irfan Munir | FAISALABAD | Urologist, Sexologist, Andrologist | MBBS, FCPS, Associate Professor of Urology | 8 | 94 | 100 | 14 | 11 |
| Assoc. Prof. Dr. Irfan Munir | FAISALABAD | Urologist, Sexologist, Andrologist | MBBS, FCPS, Associate Professor of Urology | 8 | 94 | 100 | 14 | 8 |

- **Remove Outliers from Fee(PKR) Using DBSCAN:**
  **After trying IQR & DBSCAN we found that DBSCAN was more effective at removing outliers.**
  DBSCAN removes outliers from the Fee(PKR) column by clustering data points based on density. Points that don't belong to any cluster, according to distance and minimum sample criteria, are treated as outliers and can be removed from the dataset.

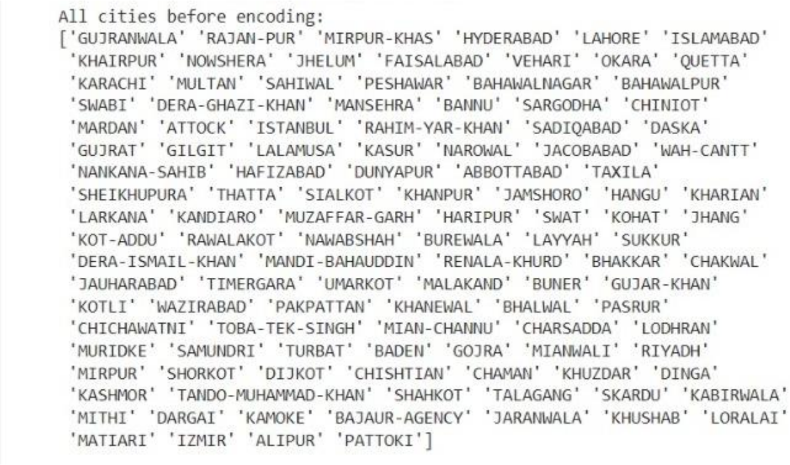


```
The DataFrame has 2190 rows before removing outliers.
The DataFrame has 2131 rows after removing outliers.
```

## 2. Columns Encoding

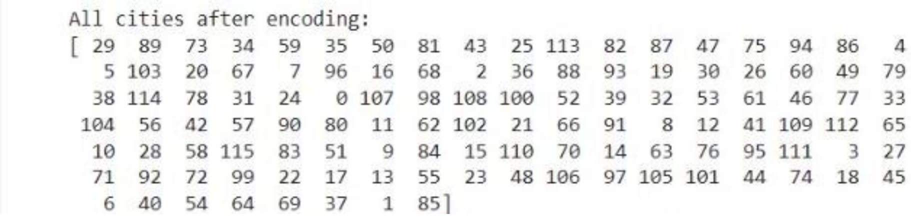
- **The 'City' column, a categorical feature, was encoded using Label Encoding. This method assigns a unique integer to each category, converting the text data into numerical format for compatibility with machine learning models. This transformation enhances model performance and training efficiency.**

**Cities Before label encoding**



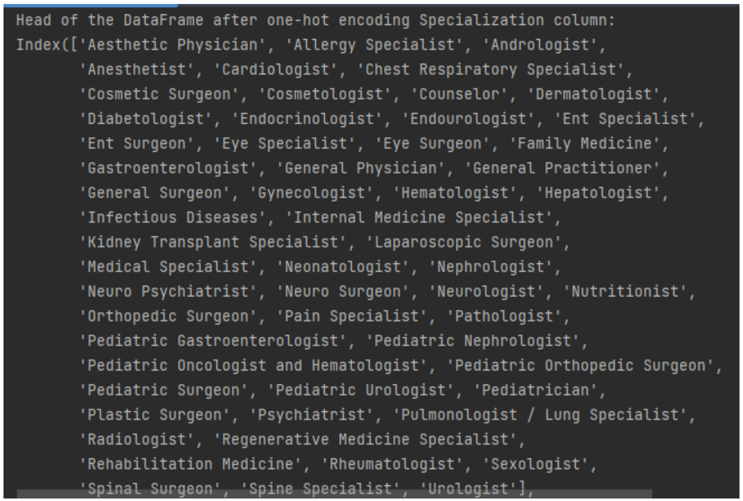
```
All cities before encoding:
['GUJRANWALA' 'RAJAN-PUR' 'MIRPUR-KHAS' 'HYDERABAD' 'LAHORE' 'ISLAMABAD'
 'KHAIRPUR' 'NOWSHERA' 'JHELUM' 'FAISALABAD' 'VEHARI' 'OKARA' 'QUETTA'
 'KARACHI' 'MULTAN' 'SAHIWAL' 'PESHAWAR' 'BAHAWALNAGAR' 'BAHAWALPUR'
 'SWABI' 'DERA-GHAZI-KHAN' 'MANSEHRA' 'BANNU' 'SARGODHA' 'CHINIOT'
 'MARDAN' 'ATTOCK' 'ISTANBUL' 'RAHIM-YAR-KHAN' 'SADIQABAD' 'DASKA'
 'GUJRAT' 'GILGIT' 'LALAMUSA' 'KASUR' 'NAROWAL' 'JACOBABAD' 'WAH-CANTT'
 'NANKANA-SAHIB' 'HAFIZABAD' 'DUNYAPUR' 'ABBOTTABAD' 'TAXILA'
 'SHEIKHUPURA' 'THATTA' 'SIALKOT' 'KHANPUR' 'JAMSHORO' 'HANGU' 'KHARIAN'
 'LARKANA' 'KANDIARO' 'MUZAFFAR-GARH' 'HARIPUR' 'SWAT' 'KOHAT' 'JHANG'
 'KOT-ADDU' 'RAWALAKOT' 'NAWABSHAH' 'BUREWALA' 'LAYYAH' 'SUKKUR'
 'DERA-ISMAIL-KHAN' 'MANDI-BAHAUDDIN' 'RENALA-KHURD' 'BHAKKAR' 'CHAKWAL'
 'JAUHARABAD' 'TIMERGARA' 'UMARKOT' 'MALAKAND' 'BUNER' 'GUJAR-KHAN'
 'KOTLI' 'WAZIRABAD' 'PAKPATTAN' 'KHANEWAL' 'BHALWAL' 'PASRUR'
 'CHICHAWATNI' 'TOBA-TEK-SINGH' 'MIAN-CHANNU' 'CHARSADDA' 'LODHRAN'
 'MURIDKE' 'SAMUNDRI' 'TURBAT' 'BADEN' 'GOJRA' 'MIANWALI' 'RIYADH'
 'MIRPUR' 'SHORKOT' 'DIJKOT' 'CHISHTIAN' 'CHAMAN' 'KHUZDAR' 'DINGA'
 'KASHMOR' 'TANDO-MUHAMMAD-KHAN' 'SHAHKOT' 'TALAGANG' 'SKARDU' 'KABIRWALA'
 'MITHI' 'DARGAI' 'KAMOKE' 'BAJAUR-AGENCY' 'JARANWALA' 'KHUSHAB' 'LORALAI'
 'MATIARI' 'IZMIR' 'ALIPUR' 'PATTOKI']
```

**Each city and its unique value**

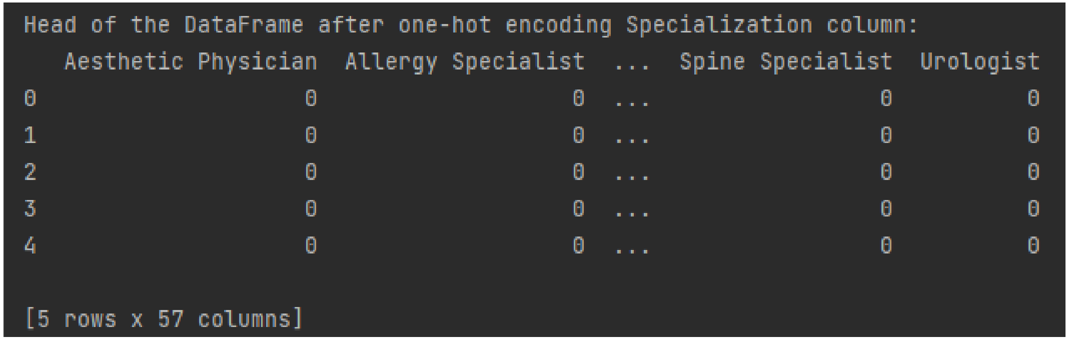


```
All cities after encoding:
[ 29  89  73  34  59  35  50  81  43  25 113  82  87  47  75  94  86   4
   5 103  20  67   7  96  16  68   2  36  88  93  19  30  26  60  49  79
  38 114  78  31  24   0 107  98 108 100  52  39  32  53  61  46  77  33
 104  56  42  57  90  80  11  62 102  21  66  91   8  12  41 109 112  65
  10  28  58 115  83  51   9  84  15 110  70  14  63  76  95 111   3  27
  71  92  72  99  22  17  13  55  23  48 106  97 105 101  44  74  18  45
   6  40  54  64  69  37   1  85]
```

- **One-hot encoding is used for the "Specialization" and "Doctor Qualification" columns, which converts each unique category into a binary column. This approach allows the model to interpret the categorical data effectively. The code also removes the original columns after encoding to avoid redundancy. This preprocessing step ensures that the model can work with the data efficiently and accurately.**
  **(it Handles the case where there's more than one qualification and specialization in one cell)**
  Specialization Columns after splitting them and will replace the Specialization column in the main data frame



```
Head of the DataFrame after one-hot encoding Specialization column:
Index(['Aesthetic Physician', 'Allergy Specialist', 'Andrologist',
       'Anesthetist', 'Cardiologist', 'Chest Respiratory Specialist',
       'Cosmetic Surgeon', 'Cosmetologist', 'Counselor', 'Dermatologist',
       'Diabetologist', 'Endocrinologist', 'Endourologist', 'Ent Specialist',
       'Ent Surgeon', 'Eye Specialist', 'Eye Surgeon', 'Family Medicine',
       'Gastroenterologist', 'General Physician', 'General Practitioner',
       'General Surgeon', 'Gynecologist', 'Hematologist', 'Hepatologist',
       'Infectious Diseases', 'Internal Medicine Specialist',
       'Kidney Transplant Specialist', 'Laparoscopic Surgeon',
       'Medical Specialist', 'Neonatologist', 'Nephrologist',
       'Neuro Psychiatrist', 'Neuro Surgeon', 'Neurologist', 'Nutritionist',
       'Orthopedic Surgeon', 'Pain Specialist', 'Pathologist',
       'Pediatric Gastroenterologist', 'Pediatric Nephrologist',
       'Pediatric Oncologist and Hematologist', 'Pediatric Orthopedic Surgeon',
       'Pediatric Surgeon', 'Pediatric Urologist', 'Pediatrician',
       'Plastic Surgeon', 'Psychiatrist', 'Pulmonologist / Lung Specialist',
       'Radiologist', 'Regenerative Medicine Specialist',
       'Rehabilitation Medicine', 'Rheumatologist', 'Sexologist',
       'Spinal Surgeon', 'Spine Specialist', 'Urologist'],
```

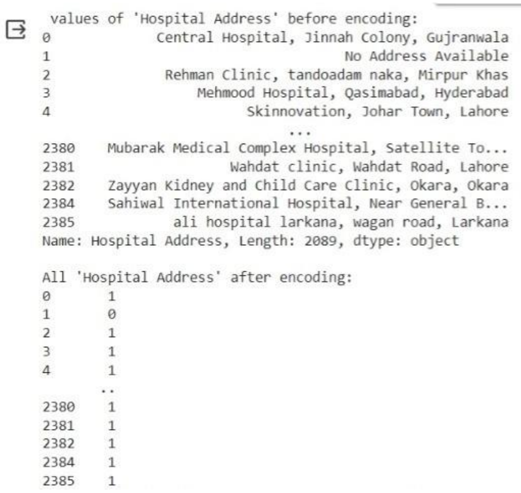


```
Head of the DataFrame after one-hot encoding Specialization column:
   Aesthetic Physician  Allergy Specialist  ...  Spine Specialist  Urologist
0                    0                   0  ...                 0          0
1                    0                   0  ...                 0          0
2                    0                   0  ...                 0          0
3                    0                   0  ...                 0          0
4                    0                   0  ...                 0          0

[5 rows x 57 columns]
```

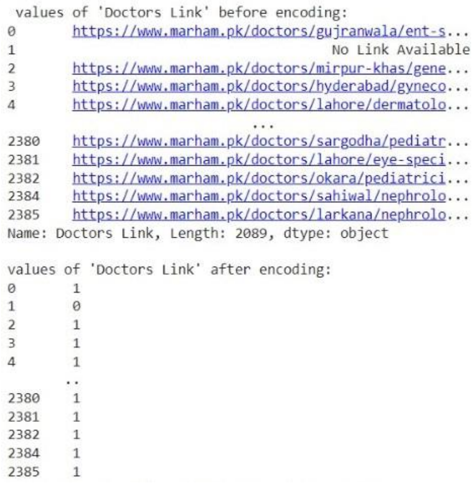
**(Similarly, with Doctor Qualification Column)**

- **Binary Encoding is used for the "Doctors Link" column, a function named “encode_doctors_link” is used to encode the data. This function transforms the data into binary values: 1 if the link starts with "https://", 0 if the link is "No Link Available", and -1 for other cases. The transformation categorizes the data into three distinct values.**
  **(Which later proved it has high correlation and usage with fee column)**



```
 values of 'Hospital Address' before encoding:
0             Central Hospital, Jinnah Colony, Gujranwala
1                                    No Address Available
2              Rehman Clinic, tandoadam naka, Mirpur Khas
3                   Mehmood Hospital, Qasimabad, Hyderabad
4                         Skinnovation, Johar Town, Lahore
                               ...
2380    Mubarak Medical Complex Hospital, Satellite To...
2381                  Wahdat clinic, Wahdat Road, Lahore
2382    Zayyan Kidney and Child Care Clinic, Okara, Okara
2384    Sahiwal International Hospital, Near General B...
2385              ali hospital larkana, wagan road, Larkana
Name: Hospital Address, Length: 2089, dtype: object

All 'Hospital Address' after encoding:
0       1
1       0
2       1
3       1
4       1
       ..
2380    1
2381    1
2382    1
2384    1
2385    1
```



```
 values of 'Doctors Link' before encoding:
0       https://www.marham.pk/doctors/gujranwala/ent-s...
1                                   No Link Available
2       https://www.marham.pk/doctors/mirpur-khas/gene...
3       https://www.marham.pk/doctors/hyderabad/gyneco...
4       https://www.marham.pk/doctors/lahore/dermatolo...
                              ...
2380    https://www.marham.pk/doctors/sargodha/pediatr...
2381    https://www.marham.pk/doctors/lahore/eye-speci...
2382    https://www.marham.pk/doctors/okara/pediatrici...
2384    https://www.marham.pk/doctors/sahiwal/nephrolo...
2385    https://www.marham.pk/doctors/larkana/nephrolo...
Name: Doctors Link, Length: 2089, dtype: object

values of 'Doctors Link' after encoding:
0       1
1       0
2       1
3       1
4       1
       ..
2380    1
2381    1
2382    1
2384    1
2385    1
```

- **Similarly, a function named “encode_hospital_address” is used to encode the "Hospital Address" column. This function transforms the data into binary values: 1 if an address is provided (not "No Address Available"), and 0 otherwise**.

## 3. Normalization

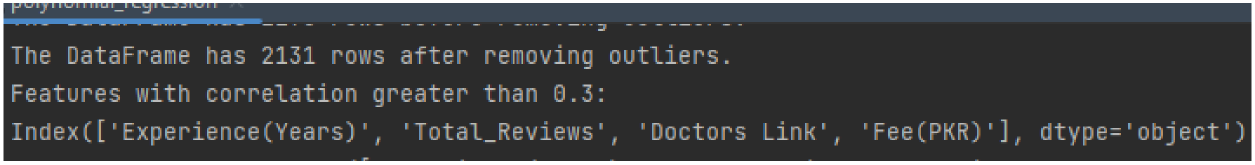
**MinMaxScaler is used to normalize all columns in the dataset except the target variable ("Fee(PKR)"). The scaler transforms the data to a range between 0 and 1 for consistency and ease of analysis.**
**We didn’t find this useful as tree-based algorithms like Random Forest and Gradient Boosting, are inherently less sensitive to the scale of the input features.**

# Feature Selection

**The code uses several methods to perform feature selection and feature engineering to identify and improve the most relevant features for predicting "Fee (PKR)" in the dataset:**

- **Filtering Features Based on Correlation:**

  **The code calculates the correlation of each feature with the target variable, "Fee(PKR)".**
  **Features with an absolute correlation value greater than 0.3 are selected.**
  **This process aims to identify the features that have a strong linear relationship with the target variable.**
  **The features that meet the correlation threshold (in this case, greater than 0.3) are printed, showing which features have the most significant relationship with the target variable.**

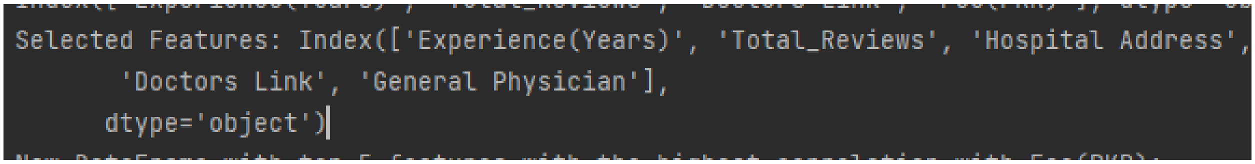


```
The DataFrame has 2131 rows after removing outliers.
Features with correlation greater than 0.3:
Index(['Experience(Years)', 'Total_Reviews', 'Doctors Link', 'Fee(PKR)'], dtype='object')
```

- **Selecting Top Features with KBest:**

  **The SelectKBest function from scikit learn is used to select the top k features based on scoring function.**

  **This method selects the top features based on their statistical relationship with the target.**



```
Selected Features: Index(['Experience(Years)', 'Total_Reviews', 'Hospital Address',
       'Doctors Link', 'General Physician'],
      dtype='object')
```

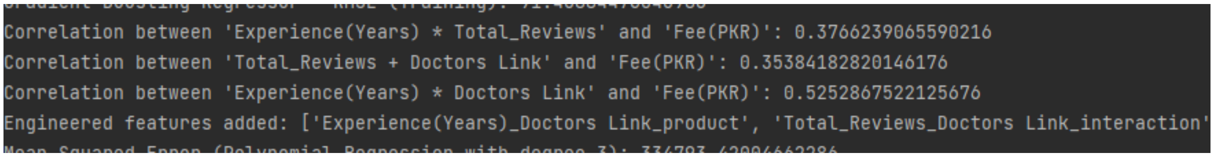
As we can see Doctors Link turned out to be important feature in relation with fee.

- ## Select from Model

  is a feature selection technique that selects important features based on a given estimator model. It ranks features according to their importance scores derived from the model and selects the ones that meet a specified threshold.

- ## Feature Engineering

  We tried to combine related features and correlated features to try and get new feature that may help us



```
Correlation between 'Experience(Years) * Total_Reviews' and 'Fee(PKR)': 0.3766239065590216
Correlation between 'Total_Reviews + Doctors Link' and 'Fee(PKR)': 0.35384182820146176
Correlation between 'Experience(Years) * Doctors Link' and 'Fee(PKR)': 0.5252867522125676
Engineered features added: ['Experience(Years)_Doctors Link_product', 'Total_Reviews_Doctors Link_interaction'
```

- ## Variance Threshold(in random forest)

  Using the VarianceThreshold method, low-variance features are removed from the dataset, allowing for more focused and efficient training on the more impactful features. This helps reduce noise and potentially improves model performance.

**Overall, these feature selection and engineering techniques aim to optimize the model by identifying and utilizing the most relevant and impactful features in the dataset**.

# Regression Models

(We used test size of 0.37 which after testing proved that it reduces overfitting and makes for better result overall)

- ## Gradient Boosting Regressor

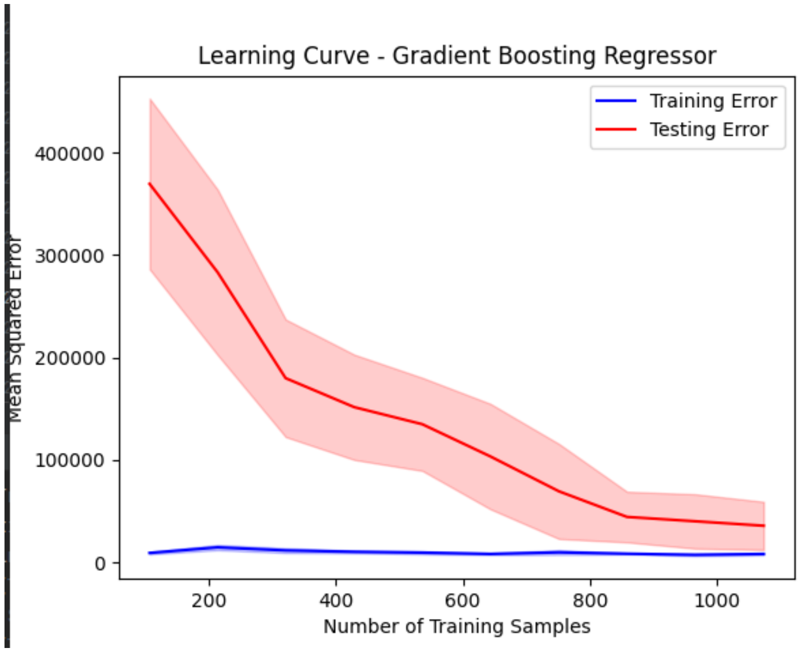
  **Gradient Boosting Regressor was chosen for its effectiveness in handling a variety of complex data sets and its ability to model non-linear relationships. GBR is an ensemble learning method that builds models sequentially and combines them to improve the overall prediction performance.**
  **We Used it with SelectFromModel after testing different feature selection techniques we found that they work best together.**

**R^2 (Testing): 0.9691237258028347**
**R^2 (Training): 0.9865179561407161**
**Mean Squared Error (Testing): 16409.470476059727**
**Mean Squared Error (Training): 8355.57690043271**
**RMSE (Testing): 128.09945540891158**
**RMSE (Training): 91.40884476040986**

**Observations Based on these results :**

The difference between training and testing MSE is low suggesting that the model is generalizing well.

The testing accuracy is close to the training accuracy this is a strong indication of good model performance.



This Graph visualizes the model's performance as we can see the test error starts high and decreases as the training set size increases. This trend indicates that the model's generalization ability improves with more data. The test error gets closer to the training error, demonstrating the model's ability to generalize better with larger datasets.

Although the test error approaches the training error, there is still a noticeable gap between them. This gap suggests that there is still room for improvement in terms of reducing the model's generalization error.

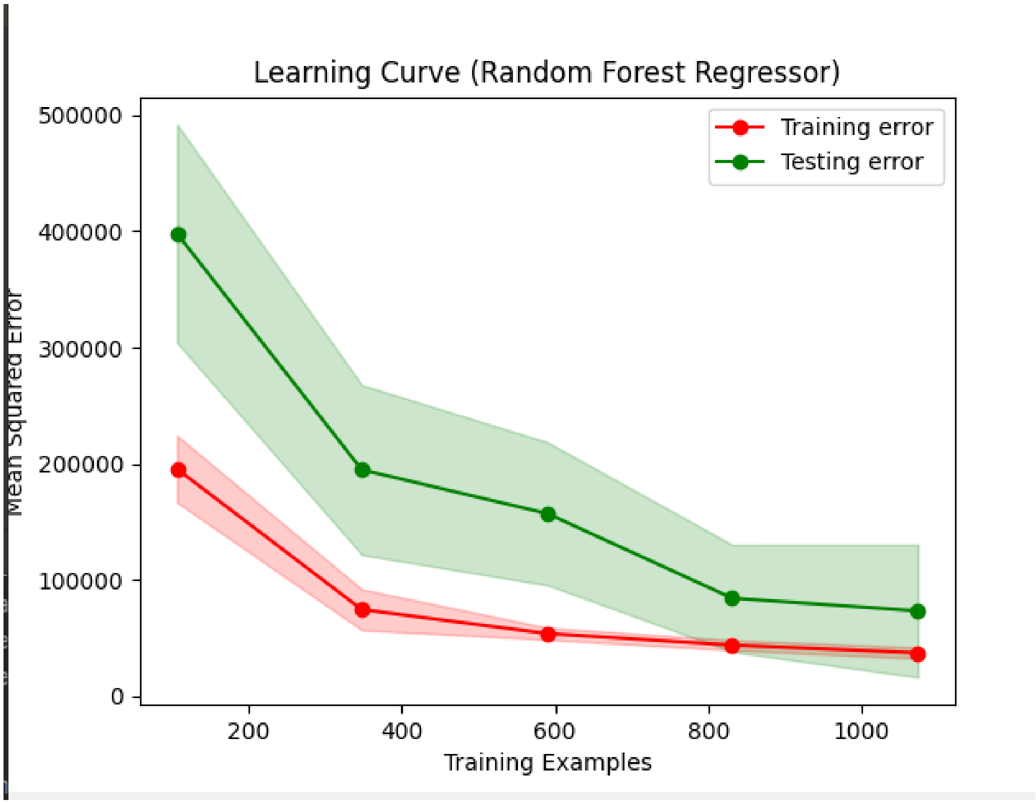
- # Random Forest

**Random Forest is an effective ensemble learning method that handles large, high-dimensional datasets well and offers feature importance insights. We used it with**

**variance threshold which removes low-variance features, streamlining the dataset for more efficient and robust model performance.**

**MSE on Testing Data: 33131.882545192326**
**MSE on Training Data: 29252.729389176155**
**RMSE on Testing Data: 182.0216540557533**
**RMSE on Training Data: 171.03429302094992**
**R^2 (Testing): 0.9376586166125163**
**R^2 (Training): 0.9527995989590845**

**Observations based on results:**
The testing data MSE is slightly higher than the training data MSE , indicating a slight difference in model performance between the datasets.
The RMSE on the testing data is also slightly higher than the training data. This is expected and suggests the model may generalize well but could benefit from re-tuning the hyperparameters.



This also indicates that there's room for improvement and the model getting better without overfitting

**Differences between random forest & gradient boost :**

Gradient Boosting trains models sequentially, each new model focusing on correcting errors of previous models. Random Forest builds multiple decision trees independently and combines their predictions.

Gradient Boosting is sensitive to hyperparameter tuning and can overfit if not properly regularized. Random Forest is generally robust with less tuning required.

- **Extra Regression Model:**
  **Polynomial Features**

**We tried to use it with different feature selection techniques (k best – highest correlated – feature engineering) and it outputted large MSE and didn’t differ much between feature techniques so it looks like its not the best regression model for our data**

Mse with feature engineering Polynomial Regression with degree : 334793.42004662286

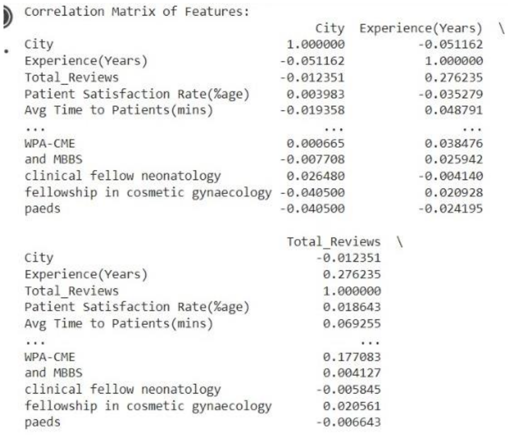
Mse with top 3 correlated feature : 329923.48411889304

Mse with Kbest : 333990.4315892749

# Extra Steps:

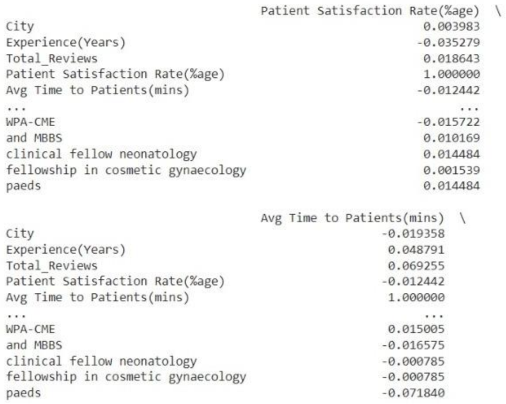
**we're using the correlation matrix to understand how different features (independent variables) are related to each other, which helps us understand the data more and therefore handle it with better techniques.**

**The correlation coefficient quantifies the strength and direction of the linear relationship between two variables. It ranges from -1 to 1**



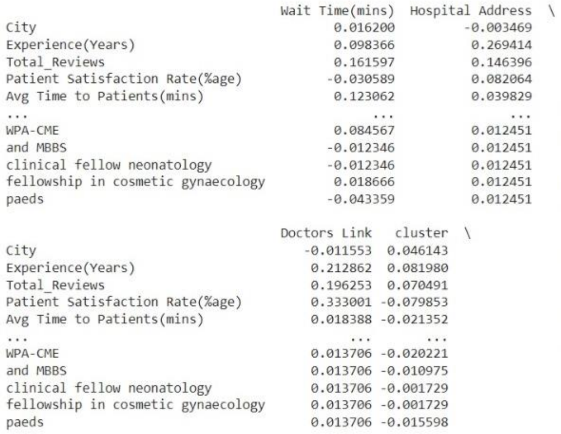
```
Correlation Matrix of Features:
                                       City  Experience(Years)  \
City                               1.000000          -0.051162
Experience(Years)                 -0.051162           1.000000
Total_Reviews                     -0.012351           0.276235
Patient Satisfaction Rate(%age)    0.003983          -0.035279
Avg Time to Patients(mins)        -0.019358           0.048791
...                                     ...                ...
WPA-CME                            0.000665           0.038476
and MBBS                          -0.007708           0.025942
clinical fellow neonatology        0.026480          -0.004140
fellowship in cosmetic gynaecology -0.040500          0.020928
paeds                             -0.040500          -0.024195

                                   Total_Reviews  \
City                                   -0.012351
Experience(Years)                       0.276235
Total_Reviews                           1.000000
Patient Satisfaction Rate(%age)         0.018643
Avg Time to Patients(mins)              0.069255
...                                          ...
WPA-CME                                 0.177083
and MBBS                                0.004127
clinical fellow neonatology            -0.005845
fellowship in cosmetic gynaecology      0.020561
paeds                                  -0.006643
```



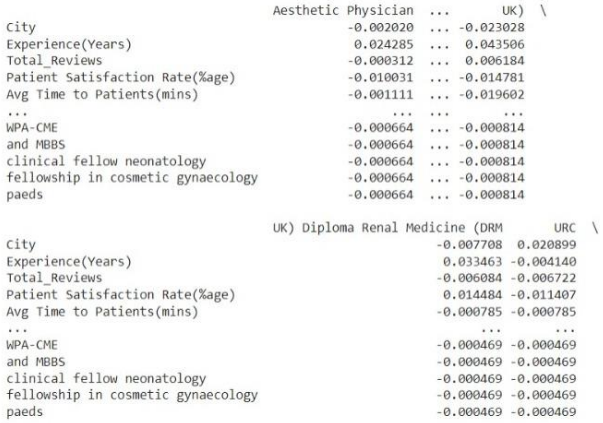
```
                                   Patient Satisfaction Rate(%age)  \
City                                                      0.003983
Experience(Years)                                        -0.035279
Total_Reviews                                             0.018643
Patient Satisfaction Rate(%age)                           1.000000
Avg Time to Patients(mins)                               -0.012442
...                                                            ...
WPA-CME                                                  -0.015722
and MBBS                                                  0.010169
clinical fellow neonatology                               0.014484
fellowship in cosmetic gynaecology                        0.001539
paeds                                                     0.014484

                                   Avg Time to Patients(mins)  \
City                                                -0.019358
Experience(Years)                                    0.048791
Total_Reviews                                        0.069255
Patient Satisfaction Rate(%age)                     -0.012442
Avg Time to Patients(mins)                           1.000000
...                                                       ...
WPA-CME                                              0.015005
and MBBS                                            -0.016575
clinical fellow neonatology                         -0.000785
fellowship in cosmetic gynaecology                  -0.000785
paeds                                               -0.071840
```



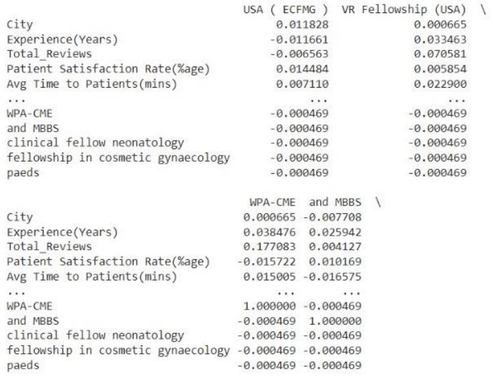
```
                                   Wait Time(mins)  Hospital Address  \
City                                      0.016200         -0.003469
Experience(Years)                         0.098366          0.269414
Total_Reviews                             0.161597          0.146396
Patient Satisfaction Rate(%age)          -0.030589          0.082064
Avg Time to Patients(mins)                0.123062          0.039829
...                                            ...               ...
WPA-CME                                   0.084567          0.012451
and MBBS                                 -0.012346          0.012451
clinical fellow neonatology              -0.012346          0.012451
fellowship in cosmetic gynaecology        0.018666          0.012451
paeds                                    -0.043359          0.012451

                                   Doctors Link   cluster  \
City                                  -0.011553  0.046143
Experience(Years)                      0.212862  0.081980
Total_Reviews                          0.196253  0.070491
Patient Satisfaction Rate(%age)        0.333001 -0.079853
Avg Time to Patients(mins)             0.018388 -0.021352
...                                         ...       ...
WPA-CME                                0.013706 -0.020221
and MBBS                               0.013706 -0.010975
clinical fellow neonatology            0.013706 -0.001729
fellowship in cosmetic gynaecology     0.013706 -0.001729
paeds                                  0.013706 -0.015598
```



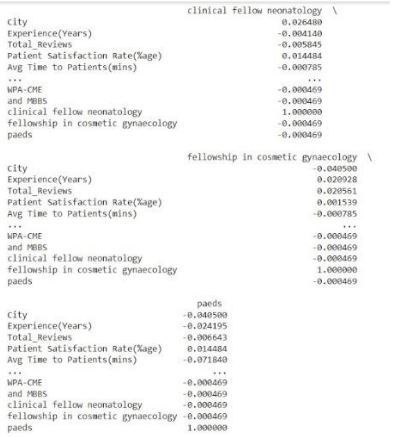
```
                                   Aesthetic Physician  ...       UK)  \
City                                         -0.002020  ... -0.023028
Experience(Years)                             0.024285  ...  0.043506
Total_Reviews                                -0.000312  ...  0.006184
Patient Satisfaction Rate(%age)              -0.010031  ... -0.014781
Avg Time to Patients(mins)                   -0.001111  ... -0.019602
...                                                ...  ...       ...
WPA-CME                                      -0.000664  ... -0.000814
and MBBS                                     -0.000664  ... -0.000814
clinical fellow neonatology                  -0.000664  ... -0.000814
fellowship in cosmetic gynaecology           -0.000664  ... -0.000814
paeds                                        -0.000664  ... -0.000814

                                   UK) Diploma Renal Medicine (DRM       URC  \
City                                                     -0.007708  0.020899
Experience(Years)                                         0.033463 -0.004140
Total_Reviews                                            -0.006084 -0.006722
Patient Satisfaction Rate(%age)                           0.014484 -0.011407
Avg Time to Patients(mins)                               -0.000785 -0.000785
...                                                            ...       ...
WPA-CME                                                  -0.000469 -0.000469
and MBBS                                                 -0.000469 -0.000469
clinical fellow neonatology                              -0.000469 -0.000469
fellowship in cosmetic gynaecology                       -0.000469 -0.000469
paeds                                                    -0.000469 -0.000469
```



```
                                   USA ( ECFMG )  VR Fellowship (USA)  \
City                                    0.011828             0.000665
Experience(Years)                      -0.011661             0.033463
Total_Reviews                          -0.006563             0.070581
Patient Satisfaction Rate(%age)         0.014484             0.005854
Avg Time to Patients(mins)              0.007110             0.022900
...                                          ...                  ...
WPA-CME                                -0.000469            -0.000469
and MBBS                               -0.000469            -0.000469
clinical fellow neonatology            -0.000469            -0.000469
fellowship in cosmetic gynaecology     -0.000469            -0.000469
paeds                                  -0.000469            -0.000469

                                    WPA-CME  and MBBS  \
City                               0.000665 -0.007708
Experience(Years)                  0.038476  0.025942
Total_Reviews                      0.177083  0.004127
Patient Satisfaction Rate(%age)   -0.015722  0.010169
Avg Time to Patients(mins)         0.015005 -0.016575
...                                     ...       ...
WPA-CME                            1.000000 -0.000469
and MBBS                          -0.000469  1.000000
clinical fellow neonatology       -0.000469 -0.000469
fellowship in cosmetic gynaecology -0.000469 -0.000469
paeds                             -0.000469 -0.000469
```



```
                                   clinical fellow neonatology  \
City                                                  0.026480
Experience(Years)                                    -0.004140
Total_Reviews                                        -0.005845
Patient Satisfaction Rate(%age)                       0.014484
Avg Time to Patients(mins)                           -0.000785
...                                                        ...
WPA-CME                                              -0.000469
and MBBS                                             -0.000469
clinical fellow neonatology                           1.000000
fellowship in cosmetic gynaecology                   -0.000469
paeds                                                -0.000469

                                   fellowship in cosmetic gynaecology  \
City                                                        -0.040500
Experience(Years)                                            0.020928
Total_Reviews                                                0.020561
Patient Satisfaction Rate(%age)                              0.001539
Avg Time to Patients(mins)                                  -0.000785
...                                                               ...
WPA-CME                                                     -0.000469
and MBBS                                                    -0.000469
clinical fellow neonatology                                 -0.000469
fellowship in cosmetic gynaecology                           1.000000
paeds                                                       -0.000469

                                       paeds
City                               -0.040500
Experience(Years)                  -0.024195
Total_Reviews                      -0.006643
Patient Satisfaction Rate(%age)     0.014484
Avg Time to Patients(mins)         -0.071840
...                                      ...
WPA-CME                            -0.000469
and MBBS                           -0.000469
clinical fellow neonatology        -0.000469
fellowship in cosmetic gynaecology -0.000469
paeds                               1.000000
```

# • Conclusion:

Our first intuition that not every column was going to be useful Like doctor name & doctors link but we used each column and tried to benefit from it & may have an importance and not rush to remove them like (Doctor’s link) columns

We also learnt different regression models and when to use each model with different types of data.

We learnt different types of encoding and when to use each encoding type with different types of data.

Lasetly we learned about how to tune our models to achieve the best performance.

And as for the next milestone we are planning to continue tuning and experimenting with different encoding , models & pre processing to achieve better results