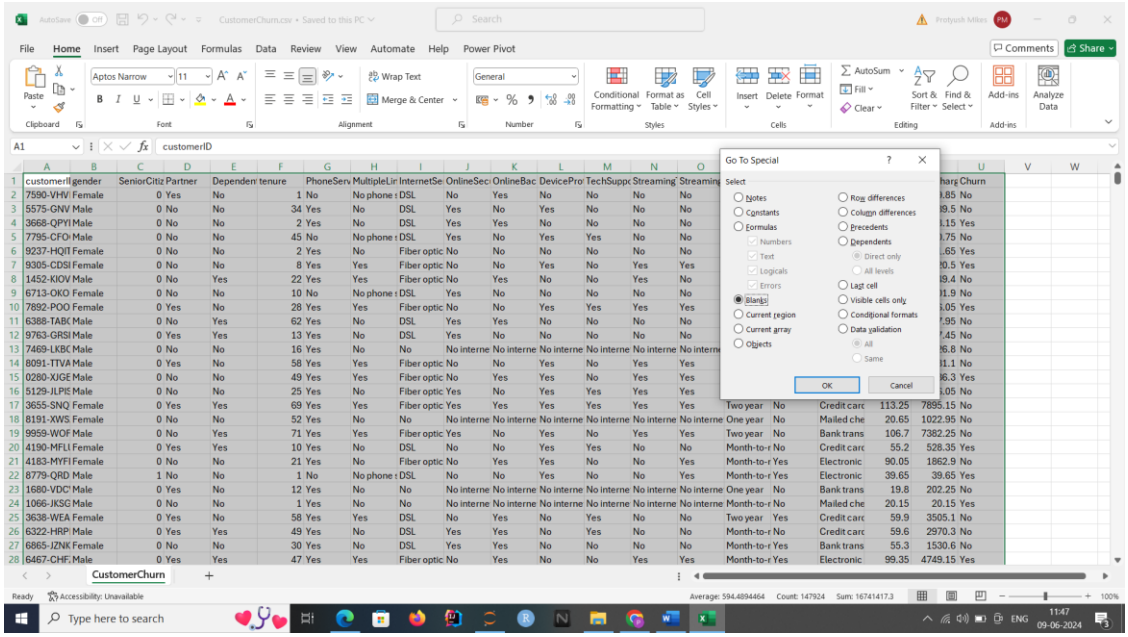
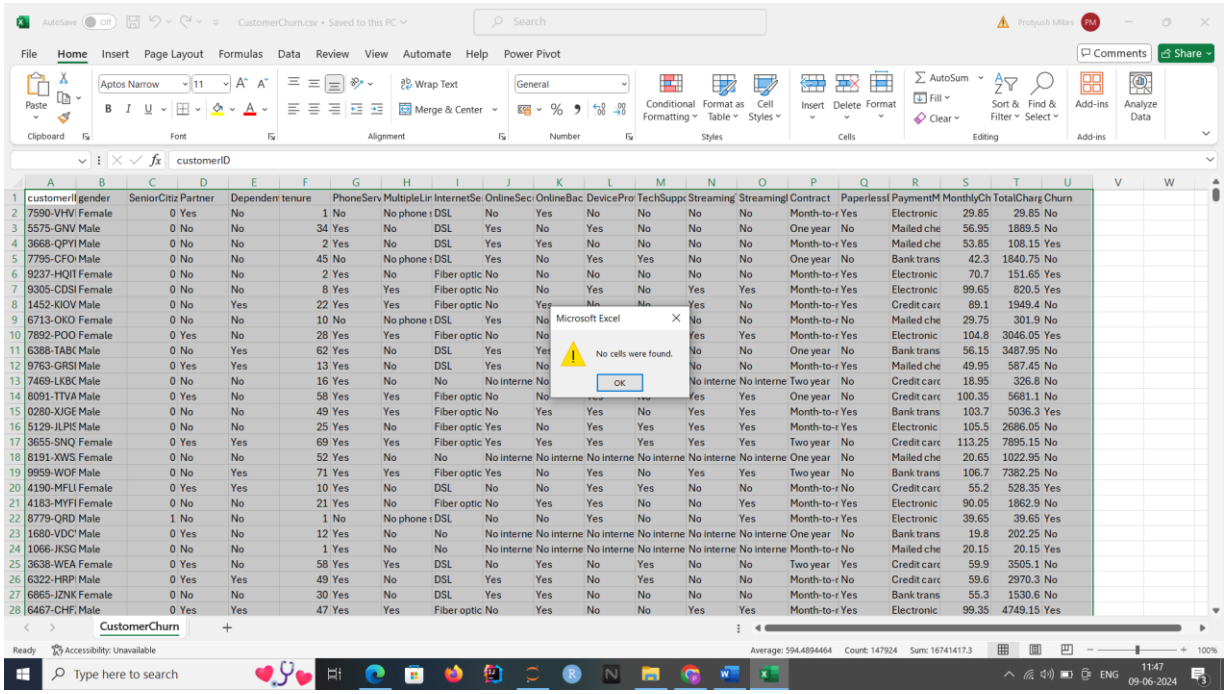
# REPORT

1. Data preprocessing: 1st our main task is to analyze the given CSV dataset in an excel file. An Excel file help us to have a look onto any missing values. Do remember that this is just a preliminary step. We will apply python code as well to see if there is any null or missing values. To do this we select all the data from the excel sheet and press F5. This will open a dialogue box and there we have to select special and then click on the blank option. It is shown in the figure below:




After clicking on blank we will get all the rows and columns highlighted that will have any missing values. Since there are no missing values on this data set , it will show no cells were found as shown below:



Now we apply Python code to see if there is any null or missing values. The 1st cell of the Jupyter notebook is there with results and it shows that there are no specific missing values. The output of the code after running the cell is provided below:

Missing values in each column:

| | |
|---|---|
| customerID | 0 |
| gender | 0 |
| SeniorCitizen | 0 |
| Partner | 0 |
| Dependents | 0 |
| tenure | 0 |
| PhoneService | 0 |
| MultipleLines | 0 |
| InternetService | 0 |
| OnlineSecurity | 0 |
| OnlineBackup | 0 |
| DeviceProtection | 0 |
| TechSupport | 0 |
| StreamingTV | 0 |
| StreamingMovies | 0 |
| Contract | 0 |
| PaperlessBilling | 0 |
| PaymentMethod | 0 |
| MonthlyCharges | 0 |
| TotalCharges | 0 |
| Churn | 0 |

```
Total number of missing values in the dataset: 0

No missing values in the dataset.
```

# 2. Exploratory Data Analysis (EDA)

Now we proceed to the EDA part. The 2nd cell in my code is designed to load, preprocess, and visualize the 'CustomerChurn.csv' dataset. We node to import important libraries like pandas for data manipulation, matplotlib and seaborn for data visualization. We also converted the CSV file into a pandas data frame using pd.read_csv(). The next step is converting the ‘Churn’ column which is our dependent variable from categorical (‘Yes’, ‘No’) to numerical (1,0). We also need to ensure that the ‘TotalCharges’ column is numberic. Non numeric values are converted to NaN. We also filled missing values in ‘TotalCharges’ with the column’s mean value using .fillna() with .mean().

Summary statistics and observations: We manged to print histograms and boxplots to visualize the numerical data distribution and identify outliers respectively.

**Observations and Summary**

**Data Information**

- The DataFrame contains multiple columns with different data types.
- The 'Churn' column has been successfully converted to numerical values.
- The 'TotalCharges' column was initially not entirely numeric but has been cleaned and missing values were handled.

**First Few Rows**

- The first few rows provide a glimpse into the structure of the dataset and verify that data loading and preprocessing steps were successful.

**Summary Statistics**

- The describe() output provides insights into the central tendency, dispersion, and shape of the dataset's distribution for numerical features.
    - **tenure**: Indicates the number of months a customer has stayed.
    - **MonthlyCharges**: Displays the monthly charges for each customer.
    - **TotalCharges**: The total amount charged to the customer.

**Visualizations**

- **Histograms**:
    - **tenure**: Shows the distribution of customers' tenure.
    - **MonthlyCharges**: Displays the distribution of monthly charges.
    - **TotalCharges**: Illustrates the distribution of total charges.
- **Boxplots**:
    - **tenure**: Identifies the range, median, and potential outliers in customer tenure.
    - **MonthlyCharges**: Shows the spread and outliers in monthly charges.
    - **TotalCharges**: Highlights the distribution and outliers in total charges.

**Key Insights**

1. **Data Completeness**:
    - The dataset has missing values in the 'TotalCharges' column, which have been imputed with the mean.
2. **Distributions**:
    - **tenure**: Distribution indicates varied customer loyalty durations.
    - **MonthlyCharges**: Monthly charges are fairly normally distributed with some concentration around certain values.
    - **TotalCharges**: Exhibits a broad range with several high-value outliers.
3. **Outliers**:
    - The boxplots reveal the presence of outliers, particularly in 'TotalCharges', indicating that some customers have significantly higher total charges than others.

I also managed to print the heatmap of correlation between the dependent and independent variable. From the observations it is clear that we will choose tenure, monthly charges, total charges and online security PaymentMethod & TechSupport as our independent variables as they have strong correlation with the dependent variables i.e, ‘Churn’.

Another EDA that I have performed on this dataset is to compare features between churned and non-churned customers using boxplots and histograms. Cells #4 and #5 shows the graphs very clearly.

## Visualizing Numerical Features by Churn

1. **Boxplots for Numerical Features:**

    - Creates boxplots for numerical features (`tenure`, `MonthlyCharges`, `TotalCharges`) to compare the distribution between churned and non-churned customers.
    - The x-axis represents the 'Churn' status (0 for non-churned, 1 for churned).

- The y-axis represents the values of the numerical features.

## Visualizing Categorical Features by Churn

2. **Countplots for Categorical Features:**

- Creates countplots for categorical features to compare the counts between churned and non-churned customers.
- The x-axis represents the categories within each feature.
- The hue represents the 'Churn' status (0 for non-churned, 1 for churned).

**Observations and Conclusions**

**Numerical Features**

1. **Tenure by Churn**:
    - **Observation**: Non-churned customers tend to have a higher median tenure compared to churned customers. The distribution shows that many churned customers have lower tenure.
    - **Conclusion**: Longer tenure is associated with lower churn rates. Customers who have been with the company for a longer period are less likely to churn.
2. **MonthlyCharges by Churn**:
    - **Observation**: Churned customers generally have higher monthly charges compared to non-churned customers. The median monthly charges are higher for churned customers.
    - **Conclusion**: Higher monthly charges might contribute to customer churn. This could indicate that customers who pay more monthly are more likely to leave, possibly due to dissatisfaction with the cost.
3. **TotalCharges by Churn**:
    - **Observation**: The distribution of total charges between churned and non-churned customers is quite similar, but churned customers show a slightly lower median total charge. There are more outliers with high total charges among non-churned customers.
    - **Conclusion**: Total charges alone may not be a strong indicator of churn. However, there is a slight trend where customers with lower total charges are more likely to churn.

**Categorical Features**

1. **Contract Type by Churn**:
    - **Observation**: Customers with month-to-month contracts have a higher churn rate compared to those with one or two-year contracts.
    - **Conclusion**: Longer contract terms are associated with lower churn rates. Customers on month-to-month plans are more likely to churn, possibly due to the lack of commitment and flexibility to leave at any time.
2. **Online Security by Churn**:
    - **Observation**: Customers without online security services have a higher churn rate.
    - **Conclusion**: Offering online security services might reduce churn. Customers with additional security features feel more secure and may be less likely to churn.
3. **Payment Method by Churn**:

- **Observation**: Customers using electronic check as a payment method have a higher churn rate compared to those using other payment methods like credit card, bank transfer, or mailed check.
- **Conclusion**: Payment method can influence churn. Customers using electronic checks might have a less positive experience or find it less convenient, leading to higher churn.

4. **Tech Support by Churn**:
   - **Observation**: Customers without tech support services have a higher churn rate.
   - **Conclusion**: Tech support services may be an important factor in reducing churn. Customers with access to tech support are likely to have better service experiences and may be less likely to leave.

**Summary**

The visualizations highlight key differences between churned and non-churned customers. Numerical features like tenure and monthly charges show clear trends with churn, indicating that longer-tenured customers and those with lower monthly charges are less likely to churn. Categorical features like contract type, online security, payment method, and tech support also show significant associations with churn rates. These insights can help in developing targeted strategies to reduce churn, such as offering incentives for longer contracts, promoting online security and tech support services, and addressing issues with specific payment methods.

# 3. Model Designing

**Explanation of Model Designing Steps:-**

1. **Loading the Dataset**:
   - The provided CSV file containing customer churn data is loaded into a pandas DataFrame.
2. **Converting Target Variable**:
   - The 'Churn' column, which indicates whether a customer has churned or not, is converted from categorical values ('Yes' and 'No') to numerical values (1 for 'Yes' and 0 for 'No') to facilitate model training.
3. **Encoding Categorical Features**:
   - Categorical features in the dataset are encoded into numerical values using LabelEncoder. This is necessary because machine learning algorithms in scikit-learn require numerical input.
4. **Feature Selection**:
   - A subset of features deemed important for predicting churn is selected. These features include 'tenure', 'MonthlyCharges', 'TotalCharges', 'Contract', 'OnlineSecurity', 'PaymentMethod', and 'TechSupport'.
5. **Defining Features and Target**:
   - The selected features are defined as the input variables (X), and the 'Churn' column is defined as the target variable (y).
6. **Splitting the Dataset**:
   - The data is split into training and testing sets using an 80-20 split. The training set is used to train the model, and the testing set is used to evaluate the model's performance.
7. **Training the Model**:

- A Random Forest Classifier is instantiated and trained using the training data. The random forest algorithm is chosen for its robustness and ability to handle both numerical and categorical data.

8. **Making Predictions**:
   - The trained model is used to make predictions on the testing set.
9. **Evaluating the Model**:
   - The model's performance is evaluated using accuracy score and classification report. The accuracy score provides the proportion of correctly predicted instances, while the classification report provides detailed metrics including precision, recall, and F1-score for each class.

**Final Observations and Conclusions**

1. **Model Accuracy**:
   - The Random Forest model achieved an accuracy of approximately 79.91% on the test data. This means that about 79.91% of the predictions made by the model are correct.
2. **Classification Report**:
   - The classification report includes precision, recall, F1-score, and support for both classes (churned and non-churned customers). These metrics provide insights into the model's performance in distinguishing between churned and non-churned customers.
   - **Precision**: Indicates the proportion of positive identifications that were actually correct. For churned customers, it shows how many predicted churns were actual churns.
   - **Recall**: Indicates the proportion of actual positives that were correctly identified. For churned customers, it shows how many actual churns were correctly identified by the model.
   - **F1-score**: A harmonic mean of precision and recall, providing a single metric that balances the two. It is especially useful when the class distribution is imbalanced.

**Summary**

- The model shows a reasonable level of accuracy (79.91%) in predicting customer churn.
- Categorical features such as 'Contract', 'OnlineSecurity', 'PaymentMethod', and 'TechSupport', along with numerical features like 'tenure', 'MonthlyCharges', and 'TotalCharges', are influential in determining whether a customer will churn.
- The precision and recall metrics indicate that the model is fairly balanced in its predictions, though there is room for improvement in identifying churned customers more accurately.
- Overall, the Random Forest model provides a solid foundation for predicting customer churn, but further tuning and perhaps additional features or different algorithms could potentially improve its performance.

Note : I also conducted an analysis using logistic regression and found that its accuracy is comparable to that of the Random Forest classifier.