***Flight Price Prediction using sagemaker***

* 1. Sagemaker :- Amazone sagemaker is a fully managed machine learning service. with sagemaker , data scientist and developers can quickly and confidently build and train, deploy ML Models into a production-ready hosted environment. It provides a UI Experience for running ml workflows that makes Sagemaker ML tools available across multiple integrated development environments(IDEs).
* Advantages:-
  + It is fully managed service
  + it provides wide range of libraries for ML workflows
  + it provide wide range of algorithms for various problems
    - Regression
    - classification
    - object detection
    - etc
  + One-click Deployment
  + it integrate with other AWS services
  + It provides notebooks and IDE for development
    - Classic Jupiter
    - Jupiter lab
    - Sagemaker Studio
  + Detailed documentation
  + It has huge community and online support.
* AWS Sagemaker Services:-
  + IAM:- It stands for Identity & access management.
    - The AWS IAM is used for managing our account in AWS though which we can create multiple users , groups and allocate the AWS resources separately so that any misuse of our account is not able to do by users and groups
    - Sagemaker IAM provide us 3 level security:-
      * IAM GROUP
      * IAM USER
      * IAM Role
  + EC2:- it stands for elastic cloud computer. it provide us cloud based computer with having the configuration that we need to get. like we want a system having 16 rams, with any Operating system and some storage then we can get it on EC2.
  + S3:- It stands for simple storage service. In this S3 bucket we store our large or any size of files we can store these files here.
* SDK:- A Software development kit is a set of platform-specific building tools for developers. you require components like debuggers, compilers, and libraries to create code that runs on a specific platform, operating system, or programming language . SDK’s put everything you need to develop and run software in one place. Additionally, they contain resources like documentation , tutorials, and guides as well as API and frameworks for faster application development.
  + Advantages:-
    - Efficient development
    - Reduce time complexity
    - Seamless Integration
    - Reduce Costs
  + AWS provide a separate SKD for Python specially called BOTO3 .
* Data cleaning & Github:-
  + Read about:- Dictionary comprehension and Dictionary unpacking.
* EDA:- It’s the practice of exploring your dataset by utilizing visualization tools and statistical measurements to understand and extract underlying patterns and information within your dataset.
  + This gives you clear picture of the data, helps us make data-informed decisions and solve crucial problems with much fewer assumptions and more facts.
  + This step is the backbone of any Data Science project and takes up a major chunk of the project timeline.
  + Statistic is used for summarization and making inference from our data by using concept like Central Tendency , Dispersion , spread, Hypothesis testing , Plots, and graphs.
  + Strength of Association:-
    - Pearson’s Correlation:- The Pearson correlation is used to know about the relationship between two numeric values, but it assumes our values are linearly separable
    - Spearman’s Rank Correlation:- The Spearman’s Rank Correlation is used to know about the relationship between two numeric variables but it assumes that our values have non-linear separation. In Today’s world our most of data have non-linear separation, so it is good to know about this Spearman’s Rank Correlation.
    - Cramer’s V :- The Cramer’s V is used to know correlation between two categorical variables.it will take all unique categories form both variables and then it create a frequency table and it run chi-square test on this table and it used the test stats value and then we calculate the Cramer’s V values. The values is in between (0 to 1), Higher the value higher the relationship between two categorical variables.
  + Hypothesis Testing:-
    - Test for Normality:- This test is used to get information about our numeric variables are following Normal Distribution or not. Here some tests are:-
      * Shapiro-Wilk Test:-
      * Anderson-Darling Test:- This test is quite more powerful than other tests for Normality Test.
    - Test For Association:- This test is used to get the information about the correlation that we have is genuine or not. Here are the test:-
      * Pearson’s Test
      * Spearman’s Test
      * Categorical Variables:-
        + Chi-Square Test
      * Numeric - Categorical Variables:-
        + One-way ANOVA Test
        + Kruskal-Wallis Test
    - Steps Involved:-
      * Sate the Hypothesis
        + Null Hypothesis
        + Alternate Hypothesis
      * Determine significance level (alpha ~5%)
      * Determine which test to perform
      * Collect necessary data (sample)
      * Obtain Critical values
      * Compute Test Statistic (and p-value)
      * Compare:-
        + Significance level vs p-value ; or
        + Critical value vs Test Statistic
      * State conclusions:- Wheater we need to accept the null hypothesis or fail to reject the null hypothesis.
  + For High level information about outliers:-
    - Isolation Forest:- Basically, it is ensemble algorithm like RandomForest . In this , at each first model numeric features and it will randomly pick one feature then randomly pick one value in that feature and then make a split on this value and same thing work with all trees
      * if a datapoint is an outlier then that point it will be capture at the top node.
      * now these steps are performed at each node
* Missingno library :- there are some built-in functions for getting information about missing value and also it generates plots , heatmap, dendrogram etc