EDA:

1) Stats-based Analysis
2) Graph-Based Analysis
3) Pre-processing:

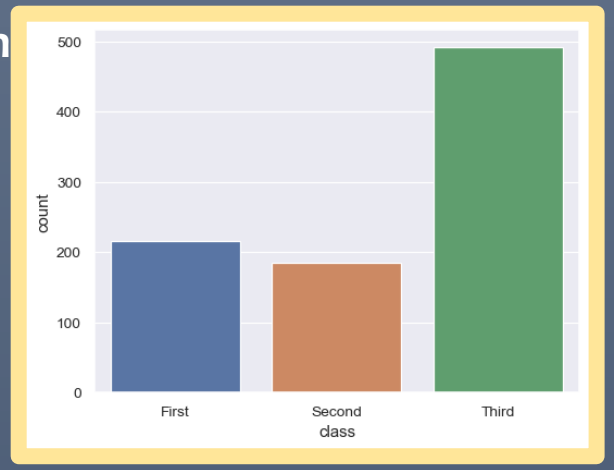
Model training

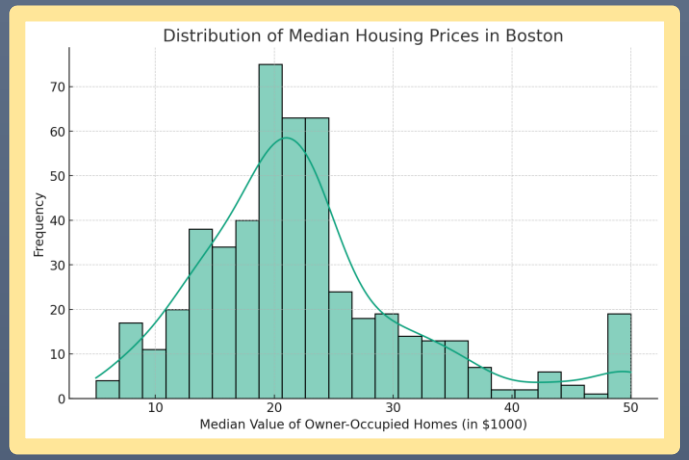
1) Training Pipeline
2) Prediction pipeline

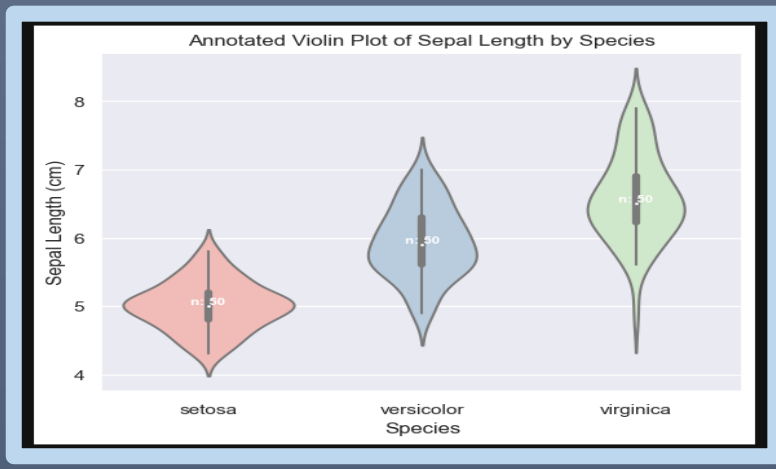
EDA:

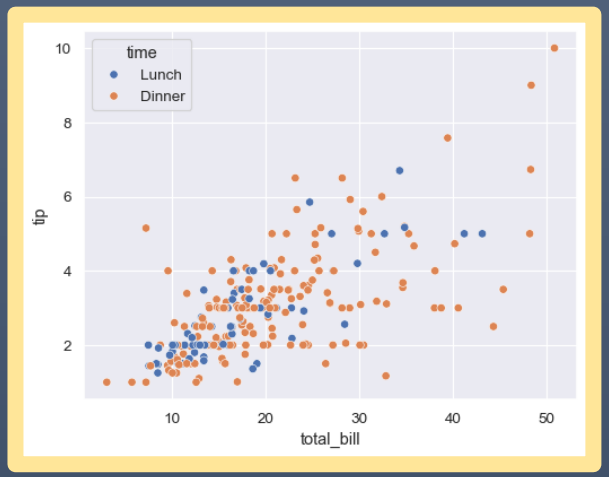
1) Stats-based Analysis
    A) Mean, median, std etc.

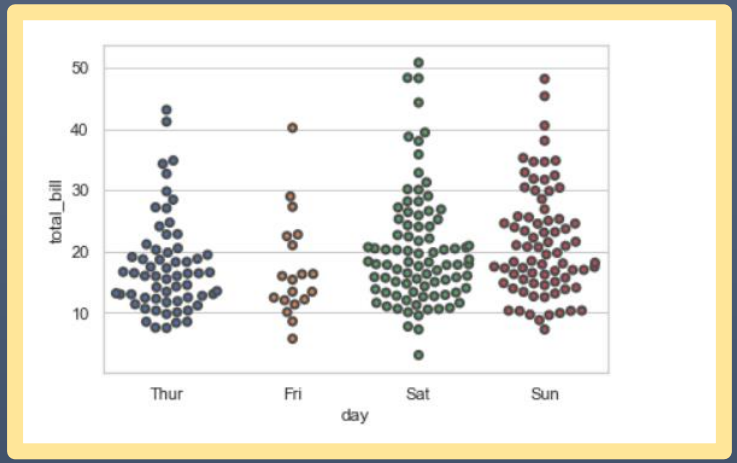
Graph

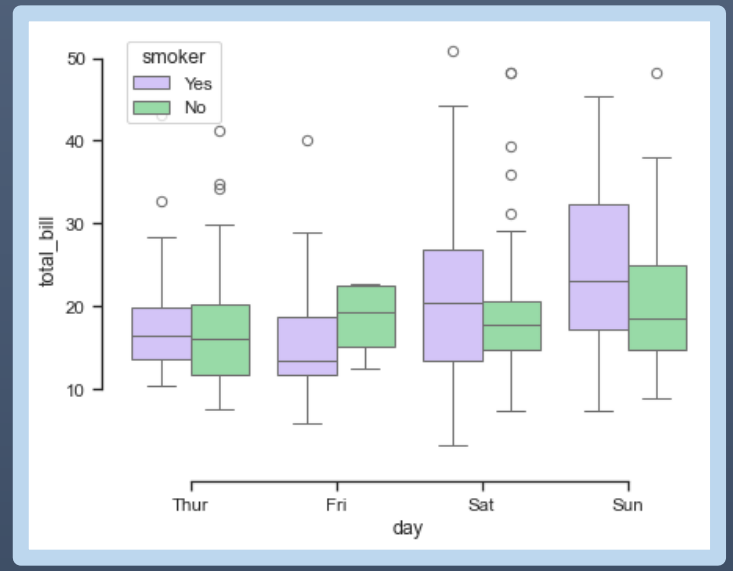

count
class
First
Second
Third

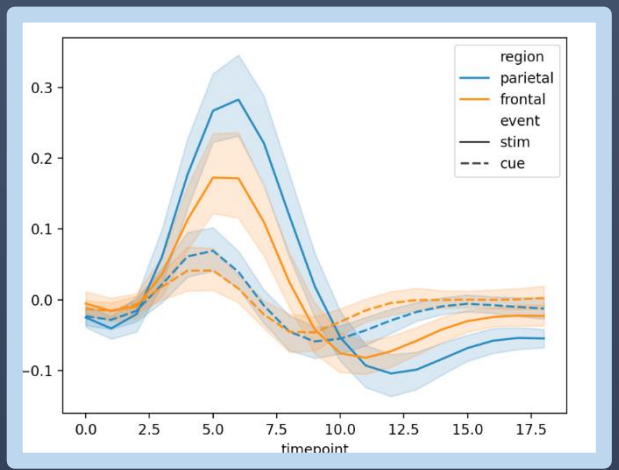

Distribution of Median Housing Prices in Boston
Frequency
Median Value of Owner-Occupied Homes (in $1000)

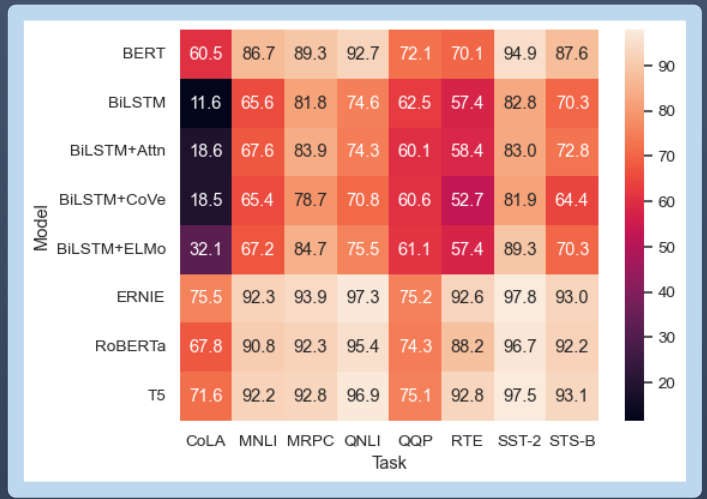

Annotated Violin Plot of Sepal Length by Species
Sepal Length (cm)
setosa
versicolor
virginica
Species


time
Lunch
Dinner
tip
total_bill


total_bill
Thur
Fri
Sat
Sun
day


smoker
Yes
No
total_bill
Thur
Fri
Sat
Sun
day


region
parietal
frontal
event
stim
cue
timepoint


BERT 60.5 86.7 89.3 92.7 72.1 70.1 94.9 87.6
BiLSTM 11.6 65.6 81.8 74.6 62.5 57.4 82.8 70.3
BiLSTM+Attn 18.6 67.6 83.9 74.3 60.1 58.4 83.0 72.8
BiLSTM+CoVe 18.5 65.4 78.7 70.8 60.6 52.7 81.9 64.4
BiLSTM+ELMo 32.1 67.2 84.7 75.5 61.1 57.4 89.3 70.3
ERNIE 75.5 92.3 93.9 97.3 75.2 92.6 97.8 93.0
RoBERTa 67.8 90.8 92.3 95.4 74.3 88.2 96.7 92.2
T5 71.6 92.2 92.8 96.9 75.1 92.8 97.5 93.1
Model
CoLA MNLI MRPC QNLI QQP RTE SST-2 STS-B
Task

Pre-processing:

1) Handle Missing Value
2) Handel Duplicate Values
3) Outlier Handling
4) Handel Imbalance Data
5) Feature Selection
6) Dimension Reduction
7) Encoding
8) Scaling

Handle Missing Value:

1) Random value
2) Forward filling/backward filling
3) Statistical approach (mean. mode, median)
4) End of the distribution
5) Create “your own” ML model to predict missing value

Handel Duplicate Values:

1) Drop the duplicate

Handling imbalance Dataset

1) Collect more data
2) Under sampling
3) Over sampling
4) Cluster based oversampling

Feature Selection

1) Correlation:
2) Variance Threshold:
3) Chi-Square Method:
4) ANOVA (Analysis of Variance):
5) Information Gain:

Dimension Reduction

1) (PCA, LDA, TSNE)
2) Split/Merge/Drop/Add

Outlier Handling:

a) Detect the Outlier
    1) z-Score
    2) IQR
    3) Boxplot
    4) Scatter plot
    5) Violin plot
b) Removing Outlier
    1) Drop Outlier
    2) Replace with other value

Encoding

1) One hot encoding
2) Ordinal Encoding

Scaling:

1) Standardization
2) Normalization (Min-max scaler)

Model training

1) Training Pipeline
   a) Choose model based on problem
   b) Hyper parameter tuning
   c) Select the best model
   d) Save model as pkl file

2) Prediction pipeline
   a) Create the API to predict the new data
   b) Handover the API to user (aws, or any other cloud)
   c) API ready for prediction