# ANALYSIS OF CAR SALE ADVERTISEMENTS

*UNVEILING TRENDS AND PATTERNS IN THE AUTOMOTIVE MARKET*

# TEAM MEMBERS - ROLES

| | |
|---|---|
| Lakshmi Prasanna Valdas | - Analyst, lead documentation |
| Dedeepya Vaddi | - Analyst, lead preprocessing |
| Sekhar Reddy Kandula | - Analyst, lead analysis/ statistical tests |
| Rajeswari Amrutha Bommaraju | - Analyst, lead statistical tests |

# MOTIVATION

- THE ANALYSIS OF CAR SALE ADVERTISEMENTS HOLDS SIGNIFICANT PROMISE FOR BOTH CONSUMERS AND INDUSTRY STAKEHOLDERS. BY DISSECTING THE WEALTH OF DATA EMBEDDED WITHIN THESE ADVERTISEMENTS, WE CAN UNLOCK INSIGHTS INTO PRICING TRENDS, CONSUMER PREFERENCES, AND THE EFFECTIVENESS OF MARKETING STRATEGIES. UNDERSTANDING THE DYNAMICS OF ONLINE CAR ADVERTISEMENTS IS ESSENTIAL IN TODAY'S DIGITAL AGE, WHERE MOST OF THE CAR BUYING JOURNEY OCCURS ONLINE. THROUGH THIS PROJECT, WE AIM TO ILLUMINATE THE FACTORS INFLUENCING CAR SALES, ENABLING BUSINESSES TO OPTIMIZE THEIR ADVERTISING EFFORTS, ENHANCE CUSTOMER SATISFACTION, AND DRIVE PROFITABILITY IN AN INCREASINGLY COMPETITIVE AUTOMOTIVE MARKET.

# DESIGN

- THE PROJECT WILL BE IMPLEMENTED USING PYTHON AND THE FOLLOWING LIBRARIES AND FRAMEWORKS:

1. **PANDAS**: A POWERFUL DATA MANIPULATION AND ANALYSIS LIBRARY FOR PYTHON. WE WILL USE PANDAS FOR LOADING THE DATASET, HANDLING MISSING VALUES, AND PERFORMING DATA PREPROCESSING TASKS.
2. **NUMPY**: A FUNDAMENTAL LIBRARY FOR SCIENTIFIC COMPUTING IN PYTHON. NUMPY WILL BE USED FOR NUMERICAL OPERATIONS AND ARRAY MANIPULATION THROUGHOUT THE PROJECT.
3. **MATPLOTLIB** AND **SEABORN**: POPULAR DATA VISUALIZATION LIBRARIES IN PYTHON. WE WILL UTILIZE THESE LIBRARIES FOR CREATING VARIOUS PLOTS AND VISUALIZATIONS DURING THE EXPLORATORY DATA ANALYSIS (EDA) PHASE.
4. **SCIKIT-LEARN**: A MACHINE LEARNING LIBRARY FOR PYTHON. SCIKIT-LEARN WILL BE USED FOR FEATURE SELECTION TECHNIQUES, BUILDING AND TRAINING DIFFERENT REGRESSION MODELS, AND EVALUATING THEIR PERFORMANCE.
5. **STATSMODELS**: A PYTHON LIBRARY FOR STATISTICAL MODELING AND DATA ANALYSIS. WE WILL EMPLOY STATSMODELS FOR CONDUCTING VARIOUS STATISTICAL TESTS, SUCH AS T-TESTS, ANOVA, AND NON-PARAMETRIC TESTS.
6. **JUPYTER NOTEBOOK**: AN OPEN-SOURCE WEB APPLICATION THAT ALLOWS US TO CREATE AND SHARE DOCUMENTS CONTAINING LIVE CODE, VISUALIZATIONS, AND NARRATIVE TEXT. WE WILL USE JUPYTER NOTEBOOK AS OUR PRIMARY DEVELOPMENT ENVIRONMENT FOR THIS PROJECT.

# DESIGN (CONTD..)

- **MODELS AND TECHNIQUES:**

1. **LINEAR REGRESSION**: AS A BASELINE MODEL, WE WILL TRAIN A LINEAR REGRESSION MODEL TO PREDICT CAR PRICES BASED ON THE SELECTED FEATURES. LINEAR REGRESSION ASSUMES A LINEAR RELATIONSHIP BETWEEN THE INDEPENDENT VARIABLES (FEATURES) AND THE DEPENDENT VARIABLE (CAR PRICE).

2. **FEATURE SELECTION TECHNIQUES**:
    - **CORRELATION ANALYSIS**: WE WILL CALCULATE CORRELATION COEFFICIENTS (E.G., PEARSON'S OR SPEARMAN'S) BETWEEN THE FEATURES AND THE TARGET VARIABLE (CAR PRICE) TO IDENTIFY HIGHLY CORRELATED FEATURES.
    - **MUTUAL INFORMATION**: WE WILL COMPUTE THE MUTUAL INFORMATION BETWEEN EACH FEATURE AND THE TARGET VARIABLE TO ASSESS THE RELEVANCE AND IMPORTANCE OF FEATURES.
    - **RECURSIVE FEATURE ELIMINATION (RFE)**: RFE IS A TECHNIQUE THAT RECURSIVELY REMOVES FEATURES WITH THE LEAST IMPORTANCE, BASED ON A SPECIFIED MACHINE LEARNING MODEL (E.G., LINEAR REGRESSION OR RANDOM FOREST).

3. **MODEL EVALUATION AND SELECTION**:
    - **CROSS-VALIDATION**: WE WILL PERFORM K-FOLD CROSS-VALIDATION TO EVALUATE THE PERFORMANCE OF OUR MODELS AND MINIMIZE THE RISK OF OVERFITTING. RELEVANT METRICS SUCH AS MEAN SQUARED ERROR (MSE), MEAN ABSOLUTE ERROR (MAE), AND R-SQUARED WILL BE COMPUTED.
    - **HOLDOUT VALIDATION**: ADDITIONALLY, WE WILL SPLIT THE DATASET INTO TRAINING AND TEST SETS AND EVALUATE THE PERFORMANCE OF OUR MODELS ON THE UNSEEN TEST SET.

4. **HYPERPARAMETER TUNING**:
    - **GRID SEARCH**: WE WILL EMPLOY GRID SEARCH TO SYSTEMATICALLY SEARCH FOR THE OPTIMAL HYPERPARAMETERS OF OUR MODELS BY EVALUATING THEIR PERFORMANCE ACROSS A SPECIFIED RANGE OF VALUES.
    - **RANDOM SEARCH**: ALTERNATIVELY, WE MAY USE RANDOM SEARCH, WHICH RANDOMLY SAMPLES HYPERPARAMETER VALUES FROM A SPECIFIED DISTRIBUTION, TO FIND THE BEST COMBINATION OF HYPERPARAMETERS.

# ABSTRACT


- TODAY, A LARGE PORTION OF THE CAR BUYER'S JOURNEY- FROM RESEARCH AND DISCOVERY TO CONDITION EVALUATION TO FINANCING AND TRANSACTION TAKES PLACE ENTIRELY ONLINE. AND A VARIETY OF FACTORS ARE AT PLAY TO PROMOTE BETTER CUSTOMER CONVERSION. THE COST OF THE CAR, VEHICLE CONDITION, MILEAGE ARE SOME FACTORS THAT RANK HIGHEST IN THE LIST OF CONSIDERATIONS.

- ONE OF THE BIGGEST OBSTACLES FOR A NEW ENTRANT IS MAINTAINING A COMPETITIVE LISTING PRICE FOR ALL THE CARS WITHOUT SACRIFICING MARGINS. THE BUSINESS NEEDS TO COMPREHEND WHAT INFLUENCES CAR PRICES IN THE MARKET. HENCE, AN IN-DEPTH ANALYSIS WILL HELP MANAGEMENT BETTER UNDERSTAND HOW PRICES CHANGE IN THE CAR-SALE MARKET.

# CONTD…

- THIS PROJECT DELVES INTO THE DYNAMIC REALM OF CAR SALE ADVERTISEMENT DATA, EMPLOYING STATISTICAL METHODOLOGIES TO UNEARTH TRENDS AND PATTERNS THAT INFLUENCE THE AUTOMOTIVE MARKET. THE PROJECT UTILIZES A COMPREHENSIVE DATASET COMPRISING DIVERSE INFORMATION ON CAR ADVERTISEMENTS, INCLUDING DEMOGRAPHIC DETAILS, PRICING, AND FEATURES. BY APPLYING INITIAL EXPLORATORY DATA ANALYSIS AND ADVANCED STATISTICAL ANALYSES, THIS ANALYSIS SEEKS TO ANSWER KEY QUESTIONS PERTAINING TO CONSUMER BEHAVIOR, MARKET DYNAMICS, AND THE EFFECTIVENESS OF MARKETING STRATEGIES IN THE AUTOMOTIVE INDUSTRY.

# GOALS

- WHICH OF THESE FACTORS- VEHICLE CONDITION, BRAND, MODEL, PRODUCTION YEAR, MILEAGE, POWER, FUEL-TYPE, CO2 EMISSION AMONG OTHERS ARE THE MOST SIGNIFICANT IN FINDING CAR PRICE IN THE MARKET?
- WHAT ARE THE AVERAGE PRICES OF CARS IN DIFFERENT REGIONS OR MARKETS?
- HOW DOES THE DEMOGRAPHIC INFORMATION OF CAR BUYERS CORRELATE WITH THE TYPES OF CARS ADVERTISED?
- IS THERE A RELATIONSHIP BETWEEN THE AGE OF A CAR AND ITS DESIRABILITY IN THE MARKET?
- HOW DO REGIONAL PREFERENCES DIFFER IN TERMS OF CAR TYPES, SIZES, OR FUEL TYPES?

# DATASET DESCRIPTION

- THE DATASET IS OBTAINED USING WEB SCRAPPING CAR ADVERTISEMENT DATA FROM OTOMOTO.PL [1] SITE

**Skoda Fabia 1.9 TDI ...**

100 HP · Diesel · 1,896 cm3 · 278,000 km · 2006 · 1.9 TDI Sportline · Fabia · Skoda

PLN 9,500

**Mazda CX-30**

149 HP · Gas · 1,998 cm3 · 8,548 km · 2022 · CX-30 · Mazda

PLN 139,900

**Mercedes-Benz CLA**

163 HP · Gas · 1,332 cm3 · 5 km · 2023 · CLA · Mercedes-Benz

PLN 193,900

**Fiat 500 1.0 GSE Hy...**

70 KM · Gas · 999 cc · 22,500 km · 2021 · 1.0 GSE Hybrid · 500 · Fiat

PLN 45,900

**Skoda Octavia 1.6 T...**

90 HP · Diesel · 1,598 cm3 · 153,832 km · 2017 · 1.6 TDI Ambition · Octavia · Skoda

PLN 45,000

**Mercedes-Benz GLB**

190 HP · Diesel · 1,950 cm3 · 5 km · 2023 · GLB · Mercedes-Benz

PLN 229,199

# CONTD…

- THE DATASET CONTAINS 25 DATA FEATURES AND 208,304 RECORDS.

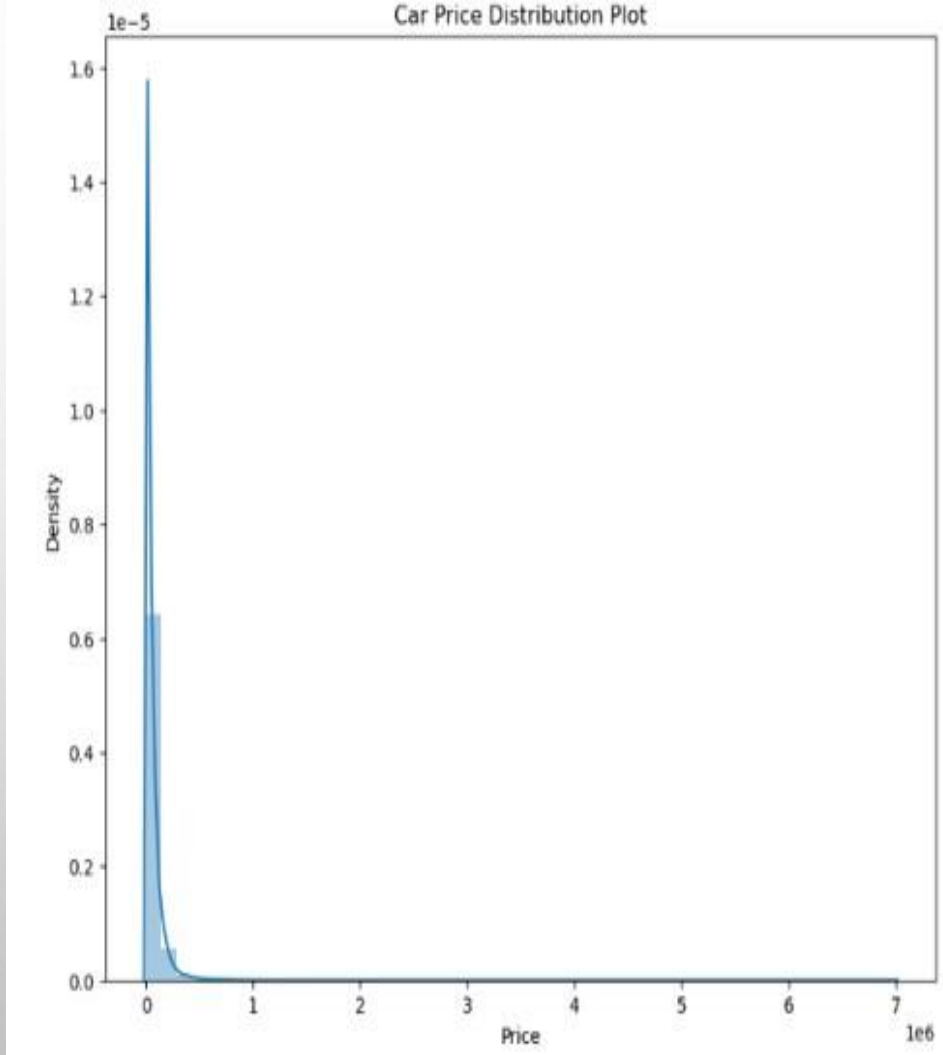
| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | Index | 208304 non-null | int64 |
| 1 | Price | 208304 non-null | int64 |
| 2 | Currency | 208304 non-null | object |
| 3 | Condition | 208304 non-null | object |
| 4 | Vehicle_brand | 208304 non-null | object |
| 5 | Vehicle_model | 208304 non-null | object |
| 6 | Vehicle_version | 138082 non-null | object |
| 7 | Vehicle_generation | 147860 non-null | object |
| 8 | Production_year | 208304 non-null | int64 |
| 9 | Mileage_km | 207321 non-null | float64 |
| 10 | Power_HP | 207661 non-null | float64 |
| 11 | Displacement_cm3 | 206338 non-null | float64 |
| 12 | Fuel_type | 208304 non-null | object |
| 13 | CO2_emissions | 94047 non-null | float64 |
| 14 | Drive | 193228 non-null | object |
| 15 | Transmission | 207825 non-null | object |
| 16 | Type | 208304 non-null | object |
| 17 | Doors_number | 206817 non-null | float64 |
| 18 | Colour | 208304 non-null | object |
| 19 | Origin_country | 118312 non-null | object |
| 20 | First_owner | 65094 non-null | object |
| 21 | First_registration_date | 86445 non-null | object |
| 22 | Offer_publication_date | 208304 non-null | object |
| 23 | Offer_location | 208304 non-null | object |
| 24 | Features | 208304 non-null | object |

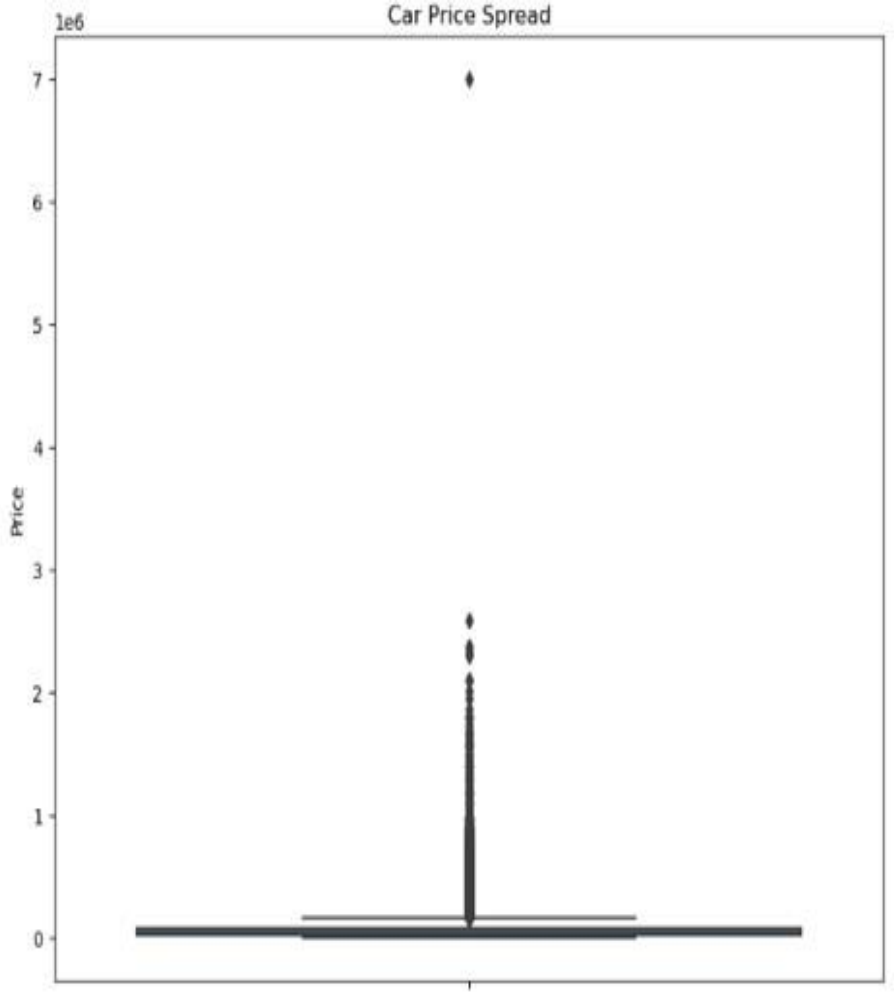
# EXPLORATORY DATA ANALYSIS

- EXPLORATORY DATA ANALYSIS (EDA) IS A CRUCIAL INITIAL STEP IN THE DATA ANALYSIS PROCESS, AIMED AT UNDERSTANDING THE STRUCTURE, PATTERNS, AND CHARACTERISTICS OF A DATASET. IT INVOLVES VISUALLY AND STATISTICALLY EXPLORING THE DATA TO UNCOVER INSIGHTS, IDENTIFY PATTERNS, DETECT ANOMALIES, AND FORMULATE HYPOTHESES FOR FURTHER ANALYSIS.

```
print(cars.Price.describe(percentiles = [0.25,0.50,0.75,0.85,0.90,1]))
```

```
count    2.083040e+05
mean     6.305383e+04
std      8.665967e+04
min      5.000000e+02
25%      1.780000e+04
50%      3.570000e+04
75%      7.599000e+04
85%      1.100000e+05
90%      1.439000e+05
100%     6.999000e+06
max      6.999000e+06
Name: Price, dtype: float64
```


Car Price Distribution Plot
1e−5
Density
Price
1e6


Car Price Spread
1e6
Price

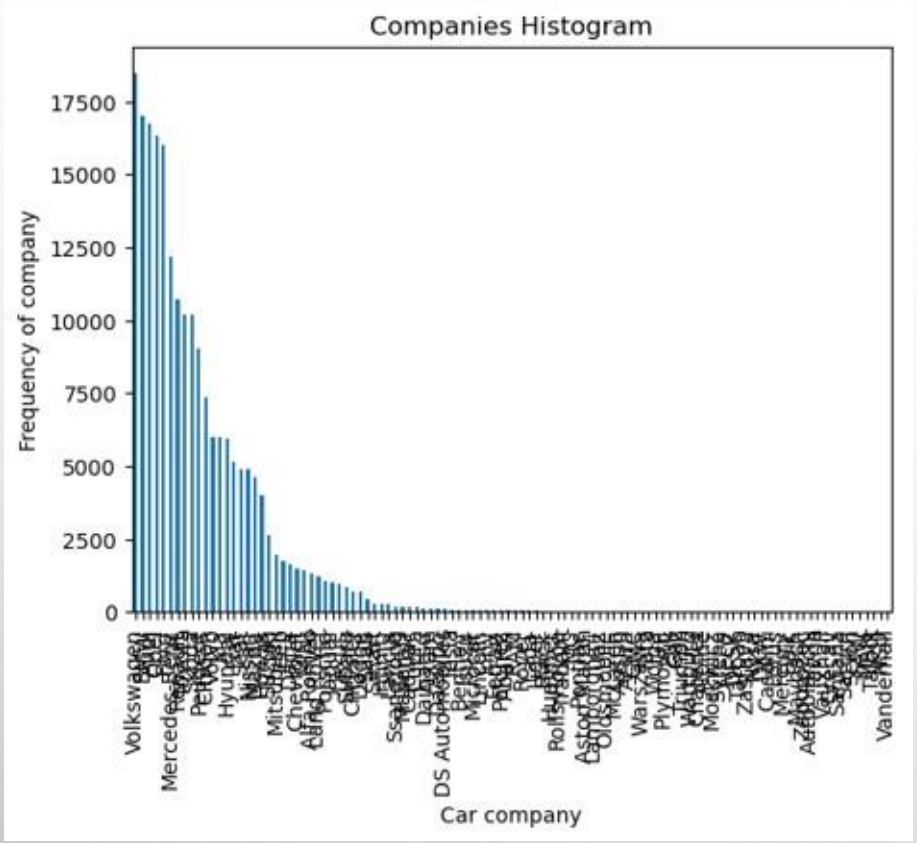



Fig. : Histogram of Frequency of Car Company

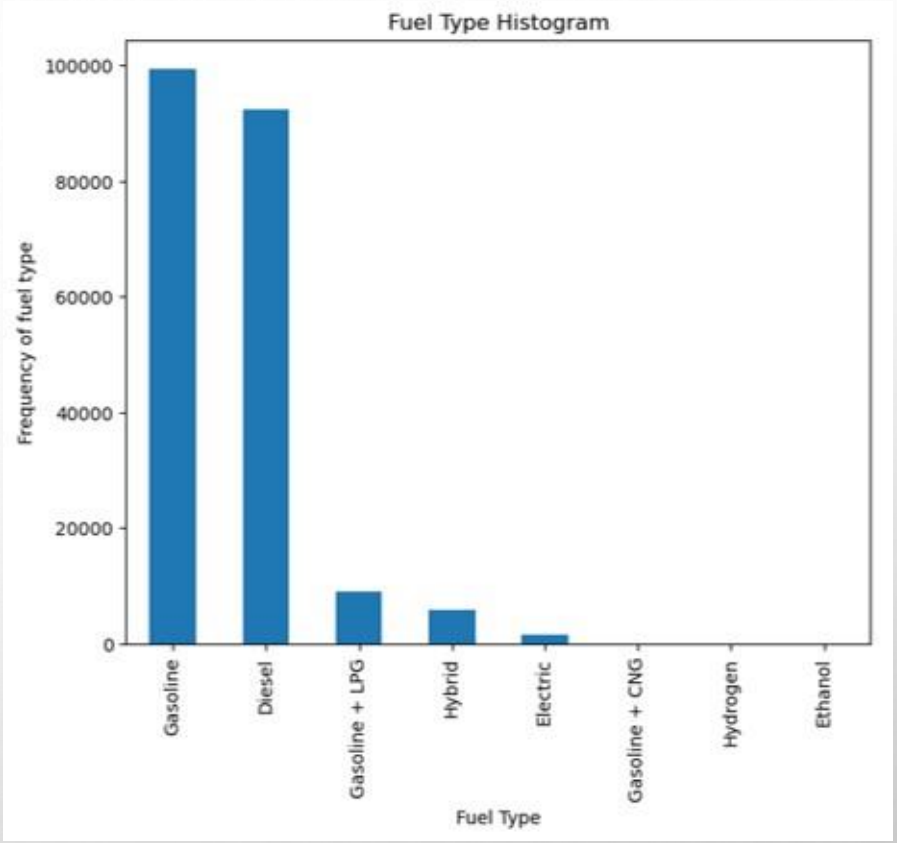



Fig. : Histogram of Frequency of Car Company

# STATISTICAL TESTS

PARAMETRIC AND NON-PARAMETRIC TESTS ARE TWO BROAD CATEGORIES OF STATISTICAL TESTS USED TO ANALYZE DATA IN DIFFERENT SCENARIOS BASED ON THE ASSUMPTIONS ABOUT THE DATA DISTRIBUTION AND THE LEVEL OF MEASUREMENT OF THE VARIABLES. HERE IS AN EXPLANATION OF PARAMETRIC AND NON-PARAMETRIC TESTS:

**PARAMETRIC TESTS**:

- **ASSUMPTION**: PARAMETRIC TESTS ASSUME THAT THE DATA FOLLOWS A SPECIFIC DISTRIBUTION, TYPICALLY A NORMAL DISTRIBUTION, AND THAT THE VARIANCES OF THE GROUPS BEING COMPARED ARE EQUAL.
- **DATA TYPE**: PARAMETRIC TESTS ARE USED FOR INTERVAL OR RATIO DATA, WHICH HAVE A KNOWN AND CONSISTENT SCALE OF MEASUREMENT.

# CONTD…

**EXAMPLES:**

- ONE-SAMPLE T-TEST: COMPARES THE MEAN OF A SINGLE SAMPLE TO A KNOWN VALUE.
- INDEPENDENT T-TEST: COMPARES THE MEANS OF TWO INDEPENDENT GROUPS.
- ANOVA (ANALYSIS OF VARIANCE): COMPARES THE MEANS OF MORE THAN TWO INDEPENDENT GROUPS.
- PEARSON CORRELATION: MEASURES THE LINEAR RELATIONSHIP BETWEEN TWO CONTINUOUS VARIABLES.

**NON-PARAMETRIC TESTS**:

- **ASSUMPTION**: NON-PARAMETRIC TESTS DO NOT MAKE ASSUMPTIONS ABOUT THE UNDERLYING DISTRIBUTION OF THE DATA AND ARE USED WHEN THE DATA IS NOT NORMALLY DISTRIBUTED OR WHEN THE VARIANCES ARE UNEQUAL.

# CONTD…

- **DATA TYPE**: NON-PARAMETRIC TESTS ARE USED FOR ORDINAL OR NOMINAL DATA, WHICH DO NOT HAVE A CONSISTENT SCALE OF MEASUREMENT.

**EXAMPLES**:

- WILCOXON SIGNED-RANK TEST: COMPARES THE MEDIAN OF A SINGLE SAMPLE TO A KNOWN VALUE.
- MANN-WHITNEY U TEST: COMPARES THE MEDIANS OF TWO INDEPENDENT GROUPS.
- KRUSKAL-WALLIS TEST: COMPARES THE MEDIANS OF MORE THAN TWO INDEPENDENT GROUPS.
- SPEARMAN CORRELATION: MEASURES THE MONOTONIC RELATIONSHIP BETWEEN TWO CONTINUOUS VARIABLES.

# SUMMARY

- ANALYSIS OF CAR SALE ADVERTISEMENTS PROJECT  OUTLINES A COMPREHENSIVE PROJECT THAT DELVES INTO THE DYNAMIC REALM OF CAR SALE ADVERTISEMENT DATA USING STATISTICAL METHODOLOGIES. THE PROJECT AIMS TO UNEARTH TRENDS AND PATTERNS INFLUENCING THE AUTOMOTIVE MARKET BY ANALYZING A DATASET COMPRISING DIVERSE INFORMATION ON CAR ADVERTISEMENTS.
- KEY ASPECTS OF THE PROJECT INCLUDE EXPLORATORY DATA ANALYSIS (EDA) TO UNDERSTAND THE DATASET, STATISTICAL TESTS SUCH AS T-TESTS, ANOVA, MANN-WHITNEY U TEST, AND KRUSKAL-WALLIS TEST FOR MEANS AND DISTRIBUTION COMPARISONS, FEATURE SELECTION USING CORRELATION ANALYSIS AND MUTUAL INFORMATION, AND MODEL SELECTION INVOLVING TRAINING, HYPERPARAMETER TUNING, AND EVALUATION.

# CONTD…

- THE IDEA EMPHASIZES THE IMPORTANCE OF UNDERSTANDING FACTORS INFLUENCING CAR PRICES, MAINTAINING COMPETITIVE LISTING PRICES, AND ANALYZING CONSUMER BEHAVIOR, MARKET DYNAMICS, AND MARKETING STRATEGIES IN THE AUTOMOTIVE INDUSTRY. THE PROJECT'S FOCUS ON BOTH PARAMETRIC AND NON-PARAMETRIC TESTS HIGHLIGHTS THE NEED FOR A COMPREHENSIVE ANALYSIS APPROACH TO DERIVE MEANINGFUL INSIGHTS FROM THE DATA.
- OVERALL, THE PROJECT DEMONSTRATES A STRUCTURED AND ANALYTICAL APPROACH TO EXPLORING CAR SALE ADVERTISEMENT DATA, AIMING TO PROVIDE VALUABLE INSIGHTS FOR MANAGEMENT TO BETTER UNDERSTAND MARKET TRENDS AND MAKE INFORMED DECISIONS IN THE AUTOMOTIVE INDUSTRY.

# FUTURE SCOPE

- FROM THE PERSPECTIVE OF THIS PROJECT ON THE ANALYSIS OF CAR SALE ADVERTISEMENTS, THERE ARE SEVERAL AVENUES FOR FURTHER EXPLORATION AND ANALYSIS THAT INDIVIDUALS OR TEAMS COULD CONSIDER:
- **SENTIMENT ANALYSIS**: INCORPORATING SENTIMENT ANALYSIS OF CAR ADVERTISEMENTS OR CUSTOMER REVIEWS COULD HELP GAUGE CONSUMER PERCEPTIONS AND SENTIMENTS TOWARDS DIFFERENT CAR MODELS OR BRANDS.
- **DYNAMIC PRICING STRATEGIES**: IMPLEMENTING DYNAMIC PRICING MODELS BASED ON MARKET DEMAND, COMPETITOR PRICING, AND OTHER FACTORS COULD OPTIMIZE PRICING STRATEGIES FOR CAR SALES.

# THANK YOU