Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

Summary of the data set

DATA CLEANING

SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

MISSING DATA

ANY MISSING DATA? HOW DID YOU HANDLE IT

DATA STORIES AND VISUALIZATIONS

THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA

ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

THIS REPORT WAS WRITTEN BY : JUSTIN BAINBRIDGE

Exploratory Data Analysis

Automobiles_csv

DATA CLEANING

Units and categorical data types are standardised and clean.

Numerous scatter plots were created based on various columns within the dataset to visually search for outliers within the set, the plots generated are realistically distributed and it is therefore assumed that there are no outliers.

When considering the Categorical data columns, the unique() function was run across the set to find any unusual or incorrect types. The results showed expected variations with no intuitive errors.

MISSING DATA

Degree of missingness:

- Isnull() was used to return the total number of NaN records present within the data set
- Initially, the "?" characters were not registered as NaN records, it was therefore necessary to replace the "?" records with np.NaN and to re-run isnull()
- A further calculation was done to determine the percentage of data that is missing, by returning the number of NaN values and dividing this value by the column size
- Overall, missing data was below 2% per column except for "normalized-"losses that shows a 20% portion of data missing. This is an uninfluential characteristic and will therefore not require these records to be removed (however analysis based on this criteria will be done with the missing values in mind).

DATA STORIES AND VISUALIZATIONS

Visualization 1

DF Column – make

Purpose – Determine which are the major manufacturers based on number of different models produced

Method – Group the data set by “make” and generate a bar plot showing sum() of amount of records per “make”.

Findings – By far the largest variety is produced by Toyota, with Nissan and Mazda contributing the second largest variety

Visualization 2

DF Column – fuel type ; aspiration ; num of doors ; body style

Purpose – The plots give a breakdown of population composition for each variety of categorical type within a series of parameters

Method – Seaborn library is used to generate countplot() visuals which sum up the number of records per category type and present the numbers in horizontal bra plots

Findings – The results show severely skewed data for the fuel type and aspiration, whereas number of doors is closer to evenly spread. The body style is predominantly Sedan and hatchback.

Visualization 3

DF Column – drive-wheels ; horsepower

Purpose – To generate a distribution visualising the range of horsepower per drive type category

Method – A seaborn distplot() was used to separate the data based on drive type, then plotted along an x-axis representing the horsepower

Findings – 4wd vehicles generally have the lower horsepower while rear wheel drive options have a stronger distribution over a high horsepower range

Visualization 4

DF Column – make ; drive wheels

Purpose – To create a distribution plot of the price ranges for a buyer looking to purchase a 4wd vehicle from either Subaru or Toyota

Method – First step was to isolate all 4wd records from the main data set. Second step made use of a seaborn distplot() to display the number of vehicles falling across the price range

Findings – Toyota overs a higher number of 4wd models within a narrow price range and Subaru offers a larger spread tending towards the higher price range

Visualization 5

DF Column – symboling

Purpose – To determine an overall risk profile of the cars within the dataset, with the higher ranking models indicating higher risk

Method – A horizontal bar graph generated by summing the number of vehicles that fall into each symbol category to give an indication of the spread

Findings – The majority of models fall within the mid range of the plot, resulting in a normal distribution with a wider tail over the higher risk end than the lower risk end. This suggests the general market takes on moderate risk with a fair demand for higher risk models

Visualization 6

DF Column – city-mpg ; highway-mpg ; horsepower

Purpose – To determine if there is a strong relationship between horsepower and improved fuel consumption in both city and highway travel

Method – Two separate scatter plots were created to display the relationship between horsepower and the two driving conditions. A second degree/polynomial regression line was included to clearly indicate the data trend

Findings – There is a strong relationship in both plots showing that fuel consumption does in fact improve with a reduction in horsepower.

Visualization 7

DF Column – symboling ; normalized-losses

Purpose – To confirm the effect of producing a model with a higher rated symbolling ranking on the normalized losses

Method – The models were categorised according to their symbolling rank and plotted on a seaborn distplot() to show how higher ranking symbolling resulted in higher normalized costs for the model

Findings – There is a clear relationship between higher ranked symbolling and increased normalized losses as expected. In order to justify these losses we would need to see additional info relating to the manufacturing and sales costs vs profits.