

**TASK**

**Exploratory Data Analysis on the Automobile Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

This dataset is based on automobiles which are categorized in whether the car is a standard or a turbo, it also determines the risk factors related to the 3 being very risky and -2 being very safe.

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

* As part of data cleaning we checked which data is missing from the data set
* We used the replace() to replace the “?” on the dataset with NaN
* We used the matrix() to display the missing values that we found from the dataset
* We used the duplicate() to determine which data on the dataset has been duplicated in which there was 0 data duplicated
* Countplot(), plot() where used to plot our visualizations

MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

* We replaced all the data that had “?” with NaN to show that this dataset for this column is not a number.
* There was no missing data meaning there was nothing to handle

DATA STORIES AND VISUALIZATIONS

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

Judging from the below visualization the diagram shows that there is no missing values in the dataset

Graphical user interface, application

Description automatically generated

This visualization shows that there is a lot of Toyota make than any other care with 32 models made while the least is a mercury make with only 1 model meaning it might just be last in the making.

Chart

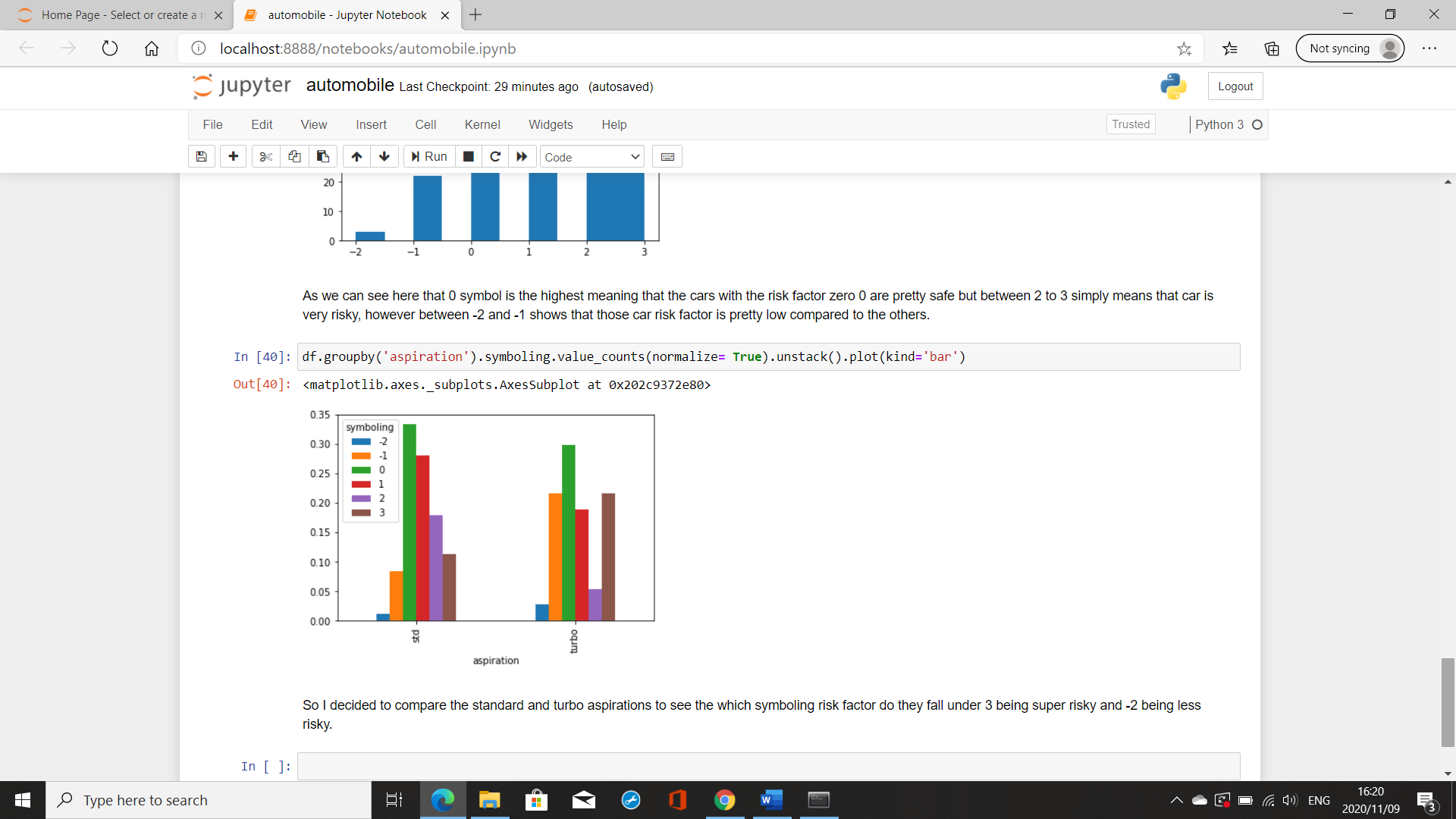
Description automatically generated

With this one we are trying to see the risk factors of the cars showing that 0 which is average safe has the highest number of care while 2 and 3 being the most risk have a high number of care which tells that people buy more care between average risk and a lot riskier. While only a few buy cars that are less risky which is between -2 and -1

Graphical user interface, text, application

Description automatically generated

Between the standard car version and the cars with the turbo we wanted to see which cars are a huge risk factors and which cars are not and from both versions we see that majority of the cars are on 0 risk factor meaning there’s not a lot of risk around that car while on turbo there is a lot of risk factor (3) around those cars than on standard cars. For pretty much safe cars (-1) we see that not a lot of cars are buying safe risk factor cars, Standard car type have less cars than turbo which are safe in terms of risk factor. Which is safe to say that people like buying cars that have a high risk factor



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

**THIS REPORT WAS WRITTEN BY : Mpho Kunene**

