Analyzing the accelerating sales of used cars in a digitally disrupted market

To cater to digitally aware used-car consumers, dealers, investors, and innovators can step up their game.

Pranav Mekal Mahesh
Computer Science & Engineering
PES University
Bangalore, India
pranavmm25@gmail.com

Mayuravarsha P
Computer Science & Engineering
PES University
Bangalore, India
mayurvp72@gmail.com

Vedant Mantri
Computer Science & Engineering
PES University
Bangalore, India
vedantmantri3@gmail.com

***Abstract*— Used-car selling is being disrupted by the internet revolution, and it's for the better. This new era of digital commerce is more than just about technology; it emphasises the importance of the consumer experience in the used-car purchase process.**

I. Introduction

Online suppliers are beginning to erode traditional used-car dealers' standing and promote growth by empowering digitally savvy customers via three primary capabilities, as indicated by our unique customer research:

- complete end-to-end purchasing capabilities
- extensive vehicle data and photos, along with effective search tools
- unique delivery options

With the development of digital players and the possibility of incumbent-dealer consolidation on the horizon, the developing market will present new dangers and opportunities for firms looking to gain a competitive advantage in an already crowded sector.

Furthermore, while customer purchasing habits are evolving, it is true that the needs of used-car buyers differ significantly from those of new-car buyers. As a result, all used-car merchants must identify their target client categories and quickly design the best ways among a growing number of accessible options in order to provide a uniformly distinctive and distinguishing customer experience.

II. CRAIGSLIST

Craigslist is a classified ads website in the United States that has sections for employment, housing, for sale, desired items, services, community service, gigs, résumés, and discussion forums.

Craig Newmark started the service in 1995 as a friend-to-friend email distribution list for local events in the San Francisco Bay Area. In 1996, it transitioned to a web-based service and expanded into more classified areas. In the year 2000, it began growing to other U.S. and Canadian cities, and it presently serves 70 countries.

Spanish, French, Italian, German, and Portuguese were the first non-English languages to be supported by Craigslist in March 2008.

With more than 49.4 million unique monthly visitors in the United States alone, the site offers more than 20 billion[12] page views each month, putting it in 72nd rank overall among websites globally and 11th place overall among websites in the United States (per Alexa.com on June 28, 2016). (per Compete.com on January 8, 2010). Craigslist is the most popular classifieds service in the world, with over 80 million new classified ads added each month.

Every month, the site receives over 2 million new job listings, making it one of the most popular job boards in the world. As of October 2011, the 23 largest U.S. cities listed on Craigslist's home page received over 300,000 listings each day in the "for sale" and "housing" sections alone. Traditional buy/sell ads, community notices, and personal ads are all included in the classified ads.

Craigslist had a workforce of 28 workers in 2009.

III. NEW PLAYERS

Carvana, Fair, and Vroom, among others, are among the new digitally savvy entrants aiming to disrupt the market. These businesses can leverage a variety of advanced digital skills, such as big data analytics and advanced digital platforms, to differentiate themselves from traditional used-car dealerships. At the same time, established new-car dealer groups and OEMs are working to maintain and increase this vital revenue stream.

The used-car inventory in the United States is getting younger and more expensive. There has been a significant movement in the market toward later-model vehicles. According to our analysis, the used-vehicle profile will grow substantially younger between 2017 and 2022, with major drop-offs in automobiles seven years and older as more consumers upgrade from older, less expensive vehicles.

IV. DATASET

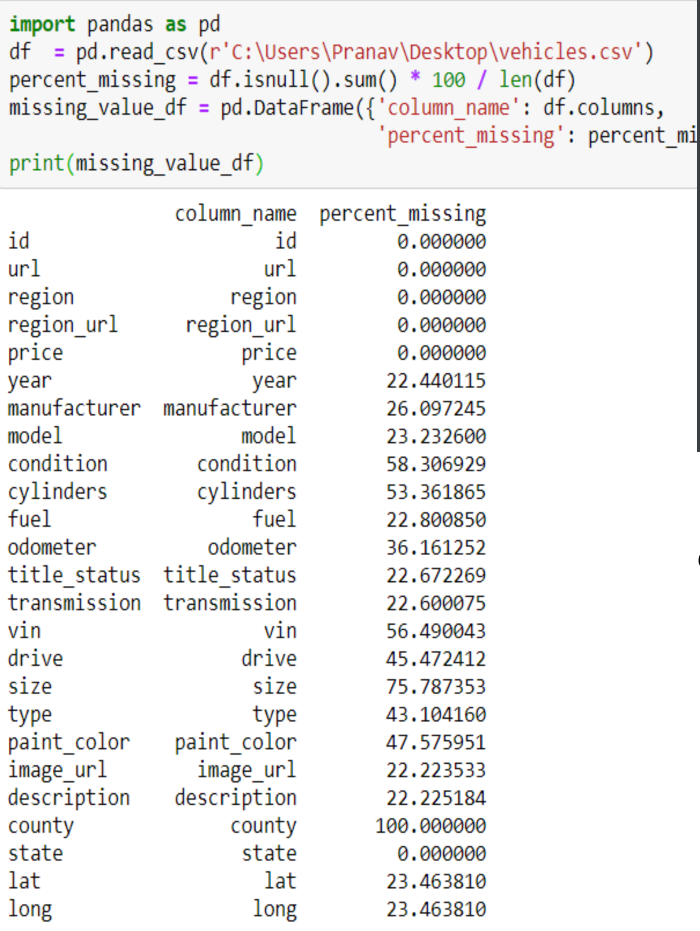
The data contains most all relevant information that Craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories.

Name of the dataset : vehicles.csv
This dataset consists of :

❑ 4
23867 rows

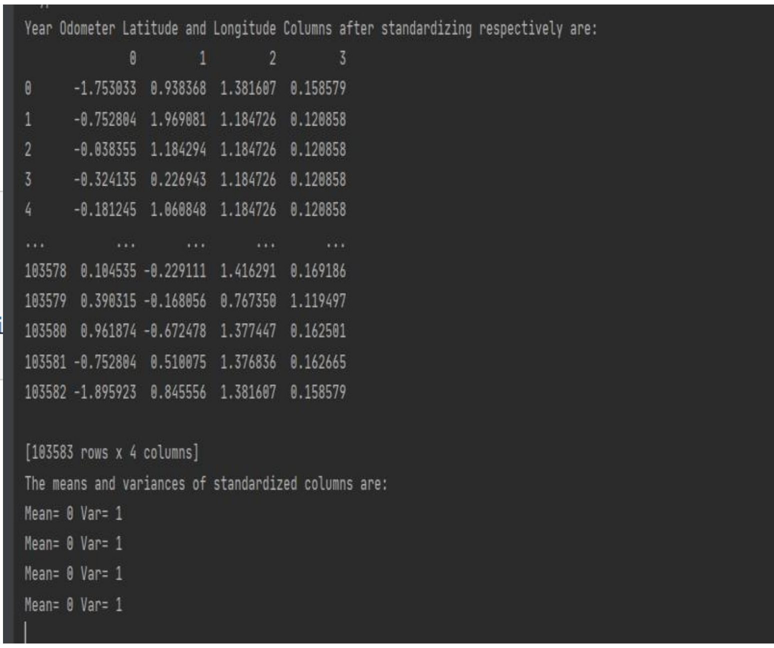
❑ 2
5 columns

❑ The image on the right shows us the percentage of missing values in each column.



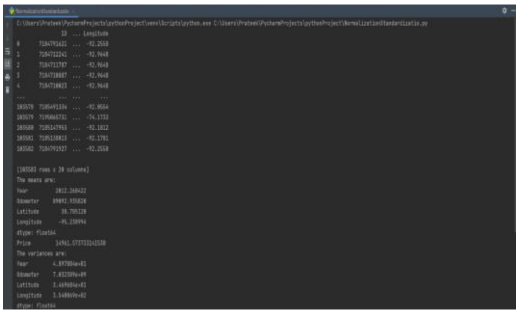
```
import pandas as pd
df  = pd.read_csv(r'C:\Users\Pranav\Desktop\vehicles.csv')
percent_missing = df.isnull().sum() * 100 / len(df)
missing_value_df = pd.DataFrame({'column_name': df.columns,
                                 'percent_missing': percent_mi

print(missing_value_df)
```

```
              column_name  percent_missing
id                     id         0.000000
url                   url         0.000000
region             region         0.000000
region_url     region_url         0.000000
price               price         0.000000
year                 year        22.440115
manufacturer  manufacturer       26.097245
model               model        23.232600
condition       condition        58.306929
cylinders       cylinders        53.361865
fuel                 fuel        22.800850
odometer         odometer        36.161252
title_status  title_status       22.672269
transmission  transmission       22.600075
vin                   vin        56.490043
drive               drive        45.472412
size                 size        75.787353
type                 type        43.104160
paint_color   paint_color        47.575951
image_url       image_url        22.223533
description   description        22.225184
county             county       100.000000
state               state         0.000000
lat                   lat        23.463810
long                 long        23.463810
```



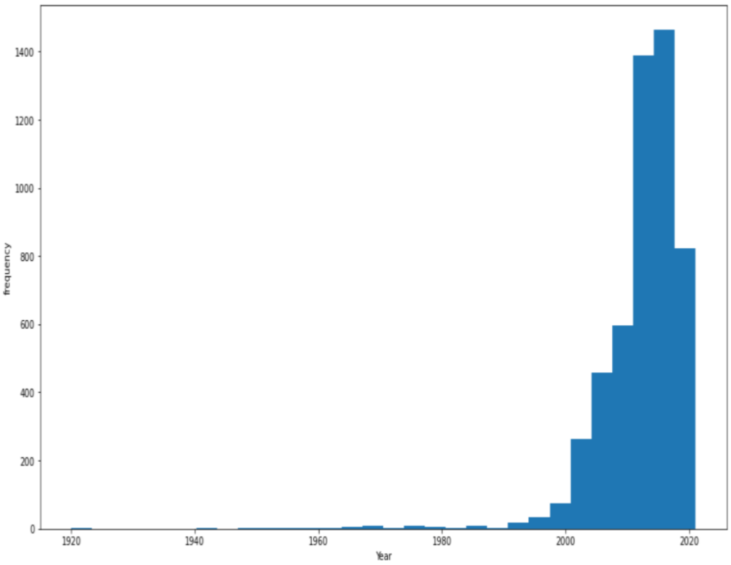
```
Year Odometer Latitude and Longitude Columns after standardizing respectively are:
              0         1         2         3
0      -1.753033  0.938368  1.381607  0.158579
1      -0.752804  1.969081  1.184726  0.120858
2      -0.038355  1.184294  1.184726  0.120858
3      -0.324135  0.226943  1.184726  0.120858
4      -0.181245  1.060848  1.184726  0.120858
...          ...       ...       ...       ...
103578  0.104535 -0.229111  1.416291  0.169186
103579  0.390315 -0.168056  0.767350  1.119497
103580  0.961874 -0.672478  1.377447  0.162501
103581 -0.752804  0.510075  1.376836  0.162665
103582 -1.895923  0.845556  1.381607  0.158579

[103583 rows x 4 columns]
The means and variances of standardized columns are:
Mean= 0 Var= 1
Mean= 0 Var= 1
Mean= 0 Var= 1
Mean= 0 Var= 1
```

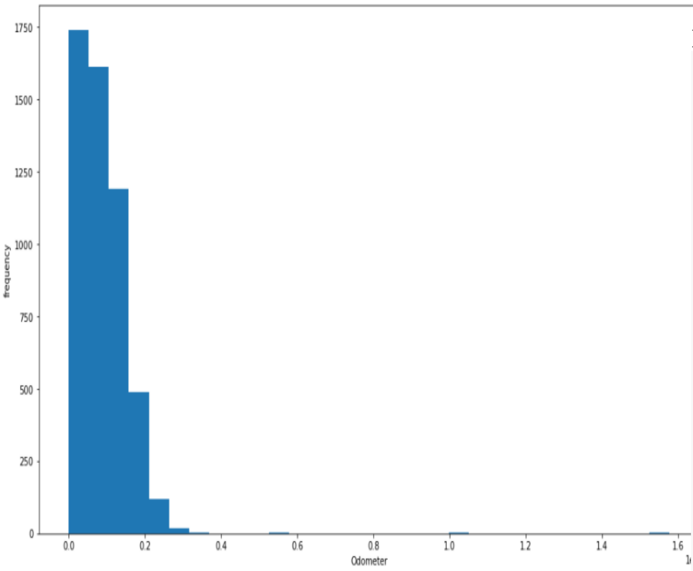
Normalization:



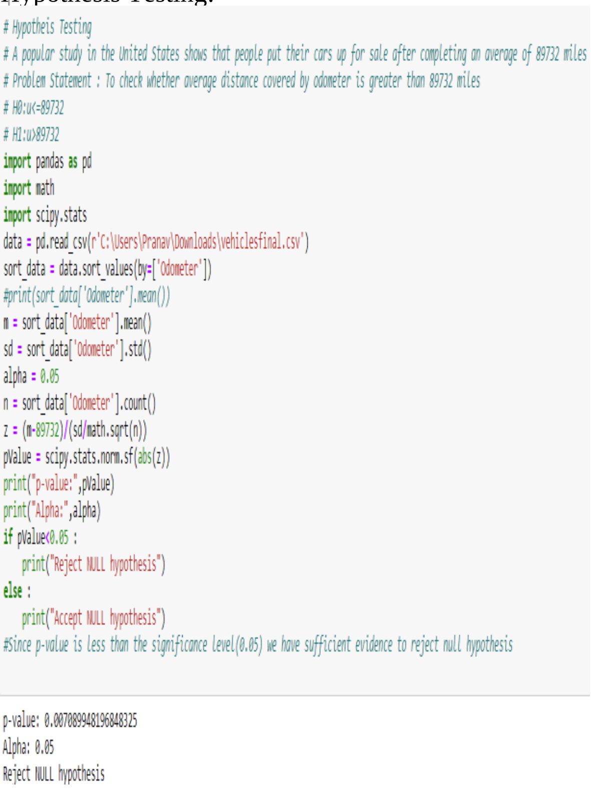
```
C:\Users\Prateek\PycharmProjects\pythonProject\venv\Scripts\python.exe C:\Users\Prateek\PycharmProjects\pythonProject\NormalizationStandardizatio.py
                ID  ...  Longitude
0       7184791621  ...   -92.2558
1       7184712241  ...   -92.9648
2       7184711787  ...   -92.9648
3       7184710887  ...   -92.9648
4       7184710023  ...   -92.9648
...            ...  ...        ...
103578  7185491334  ...   -92.0584
103579  7195665731  ...   -74.1733
103580  7185147953  ...   -92.1812
103581  7185138813  ...   -92.1781
103582  7184791927  ...   -92.2558

[103583 rows x 20 columns]
The means are:
Year          2012.268422
Odometer     89892.935828
Latitude        38.705128
Longitude      -95.258994
dtype: float64
Price         14961.573733141538
The variances are:
Year         4.897884e+01
Odometer     7.832589e+09
Latitude     3.469694e+01
Longitude    3.540869e+02
dtype: float64
```

Graphs :





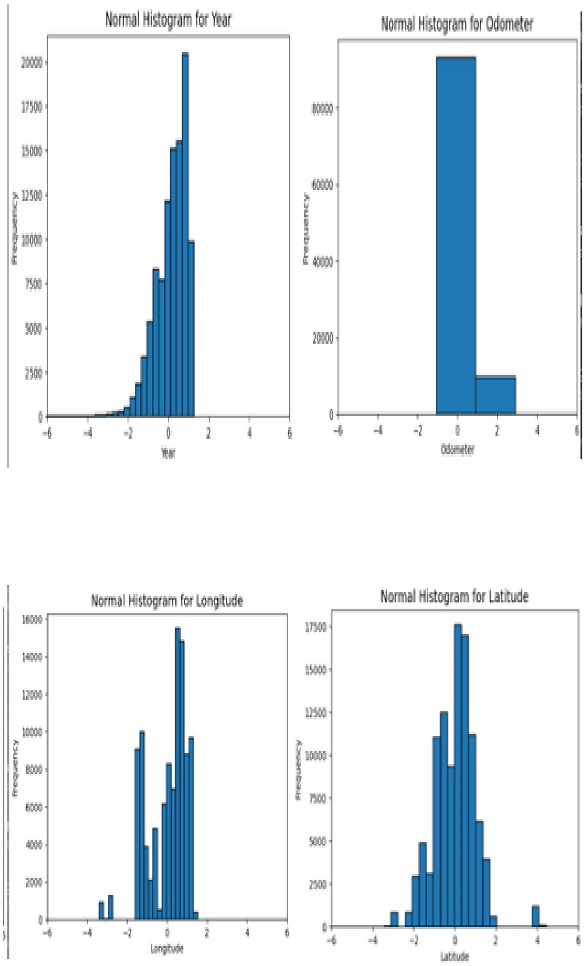
Hypothesis Testing:



```
# Hypotheis Testing
# A popular study in the United States shows that people put their cars up for sale after completing an average of 89732 miles
# Problem Statement : To check whether average distance covered by odometer is greater than 89732 miles
# H0:u<=89732
# H1:u>89732
import pandas as pd
import math
import scipy.stats
data = pd.read_csv(r'C:\Users\Pranav\Downloads\vehiclesfinal.csv')
sort_data = data.sort_values(by=['Odometer'])
#print(sort_data['Odometer'].mean())
m = sort_data['Odometer'].mean()
sd = sort_data['Odometer'].std()
alpha = 0.05
n = sort_data['Odometer'].count()
z = (m-89732)/(sd/math.sqrt(n))
pValue = scipy.stats.norm.sf(abs(z))
print("p-value:",pValue)
print("Alpha:",alpha)
if pValue<0.05 :
   print("Reject NULL hypothesis")
else :
   print("Accept NULL hypothesis")
#Since p-value is less than the significance level(0.05) we have sufficient evidence to reject null hypothesis
```

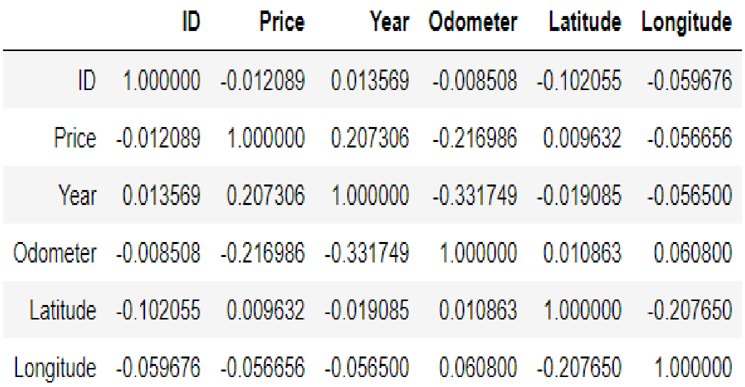
```
p-value: 0.007089948196848325
Alpha: 0.05
Reject NULL hypothesis
```

Normalization distribution graphs :



Correlation:

Here we can see that Price and Odometer are negatively related (more the distance covered, lesser the price).
Similarly Odometer and year.
Year and Price are positively related (more newer the model, higher the price).
ID, Latitude and Longitude are ignored because they have no correlation.



	ID	Price	Year	Odometer	Latitude	Longitude
ID	1.000000	-0.012089	0.013569	-0.008508	-0.102055	-0.059676
Price	-0.012089	1.000000	0.207306	-0.216986	0.009632	-0.056656
Year	0.013569	0.207306	1.000000	-0.331749	-0.019085	-0.056500
Odometer	-0.008508	-0.216986	-0.331749	1.000000	0.010863	0.060800
Latitude	-0.102055	0.009632	-0.019085	0.010863	1.000000	-0.207650
Longitude	-0.059676	-0.056656	-0.056500	0.060800	-0.207650	1.000000

V. EXISTING APPOACHES AND THEIR SHORTCOMING

A. Linear Regression

This model does not seem to work well at all as we found results where the RMSE (Root Mean Squared Error) was in the power of 19 and such models are extremely unreliable as they overfit to a huge extent and provide no value in predicting or analyzing. This model assumes that there are only linear relations in the dataset which is a huge gamble and mostly isn't necessarily true and can lead to being extremely costly as mentioned before.

B. Gradient Boosting

Does relatively well compared to Linear Regression

VI. Data Modelling

Multiple models like:

- Decision Tree Baseline
- XGBoost Baseline
- XGBoost with Parameters
- Random forest baseline
- Decision Tree HyperParam Tuning
- LightGBM with parameters
- MLP Regressor with parameter tuning
- LightGBM with categories & parameters

Are compared and analyzed to make accurate MLP models and prections

Acknowledgment

We would first like to thank our professor Dr. Gowri Srinivasa and her Teaching Assistants for guiding us throughout the course of this project. This work was supported by grants from the Computer Science and Engineering department in PES University, Bengaluru. Everyone from our team have made contributions in all aspects.

References

[1] https://www.mckinsey.com/~/media/McKinsey/Industries/Automotive%20and%20Assembly/Our%20Insights/Used%20cars%20new%20platforms%20Accelerating%20sales%20in%20a%20digitally%20disrupted%20market/Used-cars-new-platforms-Accelerating-sales-in-a-digitally-disrupted-market-vF.pdf