COMPSCI X433.3 Python for Data Analysis and Scientific Computing

Project Presentation: **Edmunds-Consumer Car Ratings and Reviews**

Edmunds-Consumer Car Ratings and Reviews

Knowing the dataset, we see there are three major car manufacturing regions with distinct car features.

Goal: Analyze the Review

1. We want to analyze if the reviews review similar car but rate them differently due to where they were manufactured?

2. We want to analyze if the reviews review rating behaviour changed with time, as new car models were better equipped with features.

3. Analyze tendency of reviewer, if they review latest cars or older cars

Context-

This is a dataset containing consumer's thought and the star rating of car manufacturer/model/type.

Content- Currently, this dataset has data of 62 major brands.

- Acura
- AlfaRomeo
- AMGeneral
- Aston Martin
- Audi
- Bentley
- BMW
- GMC
- **Toyota**
- **Volkswagen**
- honda
- Bugatti
- Buick
- Cadillac
- **Chevrolet**
- Chrysler
- Daewoo
- Dodge
- Eagle
- Ferrari
- FIAT
- Fisker
- Ford
- Genesis
- Geo
- HUMMER
- Hyundai
- INFINITI
- Isuzu
- Jaguar
- Jeep
- Kia
- Lamborghini
- Land Rover
- Lexus
- Lincoln
- Lotus
- Maserati
- Maybach
- Volvo
- Tesla
- Suzuki
- Subaru
- Spyker
- smart
- Scion
- Saturn
- Mazda
- McLaren
- Mercedes-Benz
- Mercury
- MINI
- Mitsubishi
- Nissan
- Oldsmobile
- Panoz
- Plymouth
- Pontiac
- Porsche
- Ram
- Rolls-Royce
- Saab

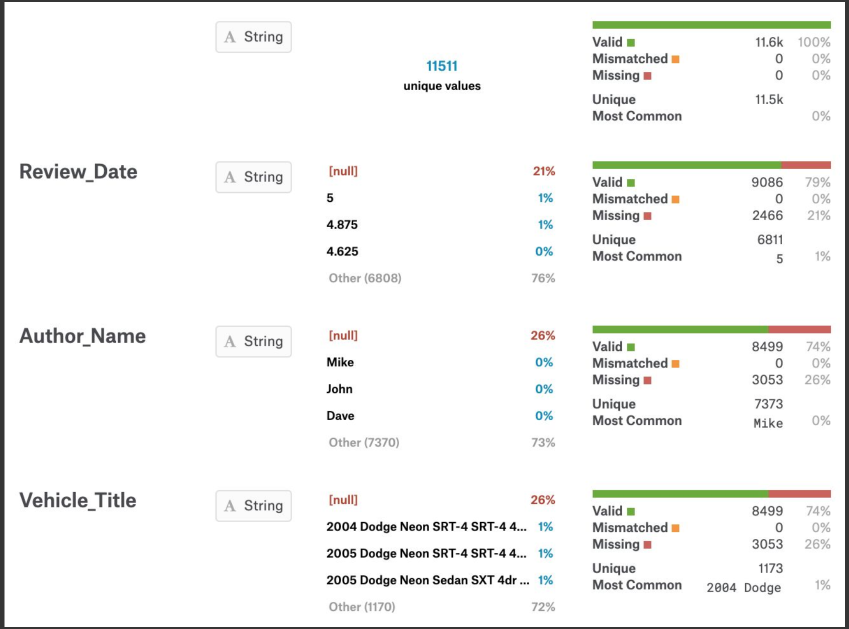
1. Dataset

Each car dataset is a separate CSV file with following columns.

- Review_Date
- Author_Name
- Vehicle_Title
- Review_Title
- Review
- Rating

Do we need all Columns for our Analysis and if the data is clean?

Column Properties



A String

11511
unique values

Valid 🟩	11.6k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	11.5k	
Most Common		0%

Review_Date — A String

[null]	21%
5	1%
4.875	1%
4.625	0%
Other (6808)	76%

Valid 🟩	9086	79%
Mismatched 🟧	0	0%
Missing 🟥	2466	21%
Unique	6811	
Most Common	5	1%

Author_Name — A String

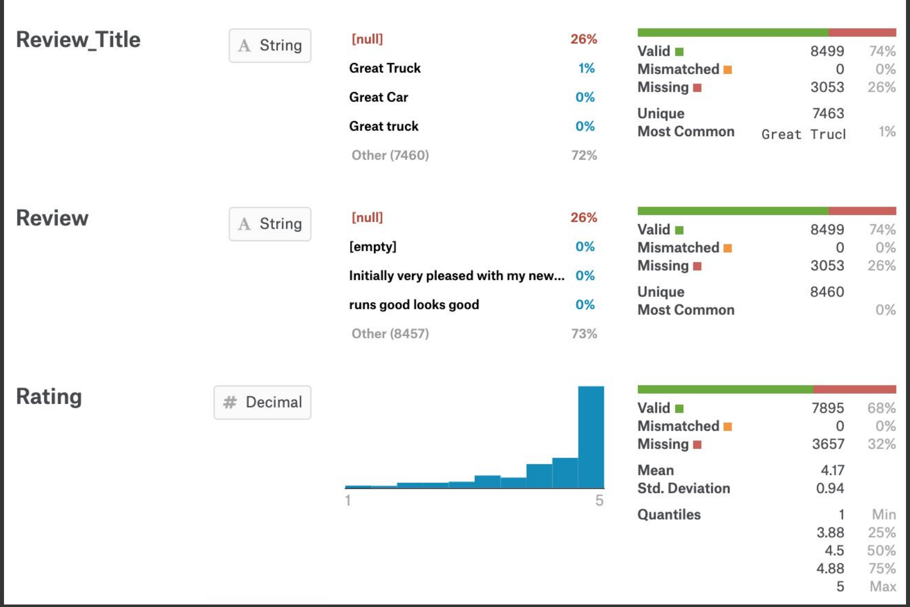
[null]	26%
Mike	0%
John	0%
Dave	0%
Other (7370)	73%

Valid 🟩	8499	74%
Mismatched 🟧	0	0%
Missing 🟥	3053	26%
Unique	7373	
Most Common	Mike	0%

Vehicle_Title — A String

[null]	26%
2004 Dodge Neon SRT-4 SRT-4 4...	1%
2005 Dodge Neon SRT-4 SRT-4 4...	1%
2005 Dodge Neon Sedan SXT 4dr ...	1%
Other (1170)	72%

Valid 🟩	8499	74%
Mismatched 🟧	0	0%
Missing 🟥	3053	26%
Unique	1173	
Most Common	2004 Dodge	1%



Review_Title — A String

[null]	26%
Great Truck	1%
Great Car	0%
Great truck	0%
Other (7460)	72%

Valid 🟩	8499	74%
Mismatched 🟧	0	0%
Missing 🟥	3053	26%
Unique	7463	
Most Common	Great Trucl	1%

Review — A String

[null]	26%
[empty]	0%
Initially very pleased with my new...	0%
runs good looks good	0%
Other (8457)	73%

Valid 🟩	8499	74%
Mismatched 🟧	0	0%
Missing 🟥	3053	26%
Unique	8460	
Most Common		0%

Rating — # Decimal

Valid 🟩	7895	68%
Mismatched 🟧	0	0%
Missing 🟥	3657	32%
Mean	4.17	
Std. Deviation	0.94	
Quantiles	1	Min
	3.88	25%
	4.5	50%
	4.88	75%
	5	Max

Dataset: Analysis and Cleanup

Review_Date:

This is a object, need to clean and transform to datetime.

Author_Name:

In order to understand the behaviours, we need to retain Authors

Vehicle_Title:

We need to extract Year of Manufacturing as separate column. Rest we can discard.

Review_Title:

We can drop the data, as we are not doing sentiment analysis

Review:

We can drop the data, as we are not doing sentiment analysis

Rating:

We will keep the data but convert them to categorical data with only ratings in whole number

Dataset: Cleaned

After cleaning dataset we are going to be left with Review_Date, Author_Name, Rating and Model_Year

What’s Next

Analyze the data quality and inconsistencies and address them

Findings:

Review_Date: Some review date were in format which cannot be converted to Datetime

Author_Name: Reviewers provided review anonymously

Rating: Rating is continuous variable with decimals, which causes too many data points

Model_Year: Some Car details were missing Model year.

Rating: Some reviewers, provided review comments but did not provide rating.

Tip

Different columns has different reason for missing data or inconsistent data. So the approach would be different.

What’s Next

How data were fixed

Review_Date:

For missing review date, we looked for Model year and populated it as Review Date and also used forward fill if Model year is missing.

Author_Name:

Reviewers provided review anonymously, so for all missing Authors, we added Author as ‘anonymous’

Rating:

Fill the missing values with mean for the rating

Convert the rating to floor value

Model_Year:

Fetch the review year and populate as Model year

Tip

For Author Review tendency, we will retain the original dataset

Analyze Author Review Behaviour for cars from Different Region

How to Prepare Data

1) Add a column for all three dataset called Region
2) Find the missing data
3) Find columns from where we can extract data of interest
4) Fill NnN appropriately
5) Append all three dataset

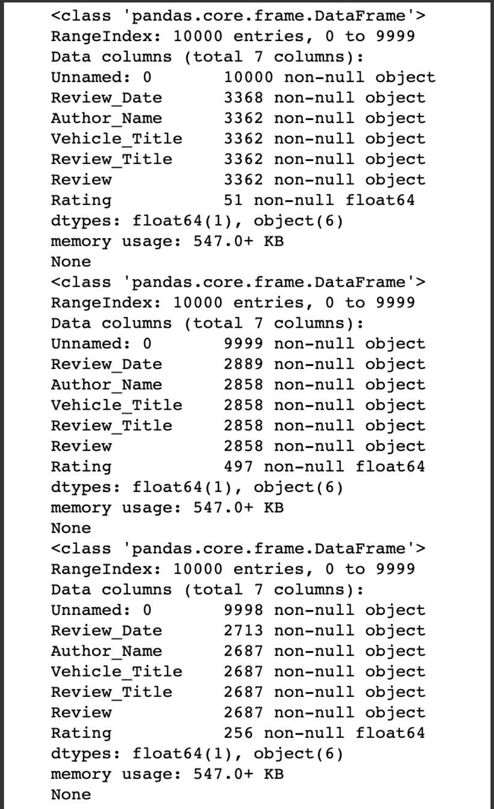
Analysis

1) Group by Reviewer, Region and Model Year
2) Draw graph depicting the distribution of rating by Model year
3) Draw graph depicting the distribution of rating by Region
4) Look top reviewed and their average rating given
5) Explore if they are consistently reviewing cars across the board

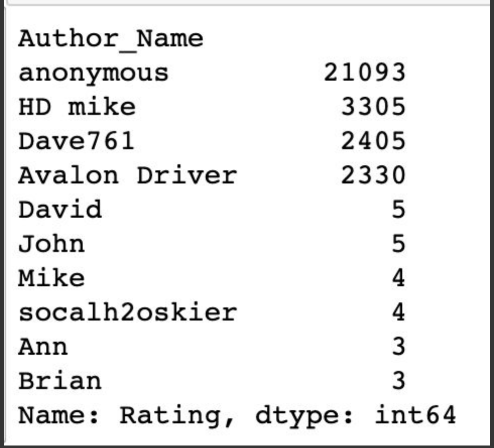
Conclusion

1) Most of the reviews were submitted anonymously
2) 99% time, reviewers rate 5 for cars. That gives an idea that happy reviewers tends to review more than unhappy
3) Cars across the region tends to get consistent reviews
4) Top 3 reviews shows come inconsistent behavior, they provided all of their rating 5 and in same year

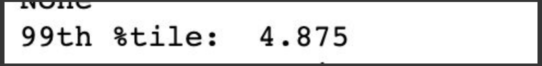
Some Outputs



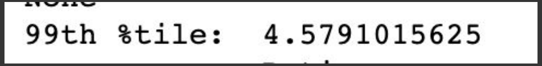
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
Unnamed: 0       10000 non-null object
Review_Date      3368 non-null object
Author_Name      3362 non-null object
Vehicle_Title    3362 non-null object
Review_Title     3362 non-null object
Review           3362 non-null object
Rating           51 non-null float64
dtypes: float64(1), object(6)
memory usage: 547.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
Unnamed: 0       9999 non-null object
Review_Date      2889 non-null object
Author_Name      2858 non-null object
Vehicle_Title    2858 non-null object
Review_Title     2858 non-null object
Review           2858 non-null object
Rating           497 non-null float64
dtypes: float64(1), object(6)
memory usage: 547.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
Unnamed: 0       9998 non-null object
Review_Date      2713 non-null object
Author_Name      2687 non-null object
Vehicle_Title    2687 non-null object
Review_Title     2687 non-null object
Review           2687 non-null object
Rating           256 non-null float64
dtypes: float64(1), object(6)
memory usage: 547.0+ KB
None
```



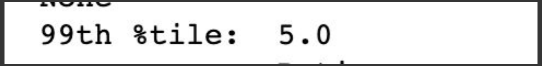
```
Author_Name
anonymous         21093
HD mike            3305
Dave761            2405
Avalon Driver      2330
David                 5
John                  5
Mike                  4
socalh2oskier         4
Ann                   3
Brian                 3
Name: Rating, dtype: int64
```



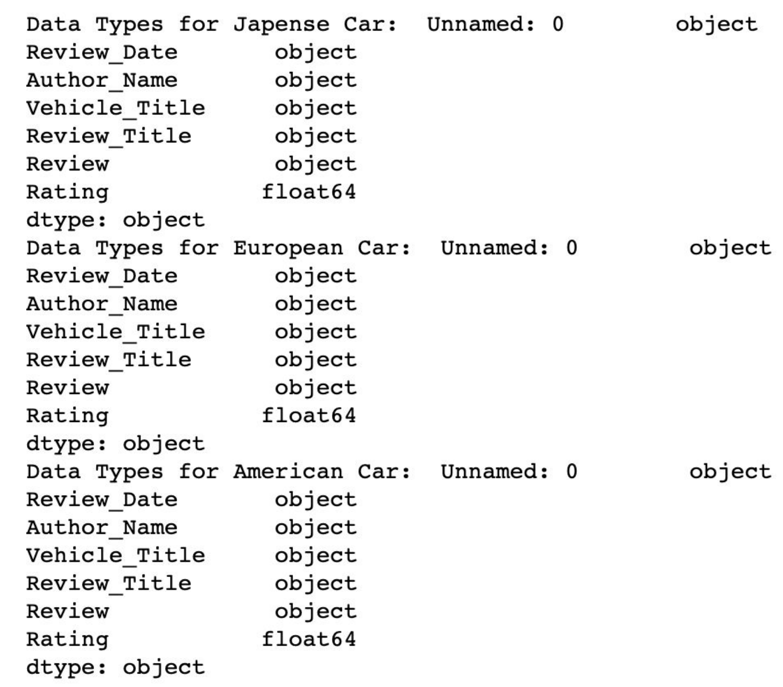
```
None
99th %tile:  4.875
```



```
None
99th %tile:  4.5791015625
```

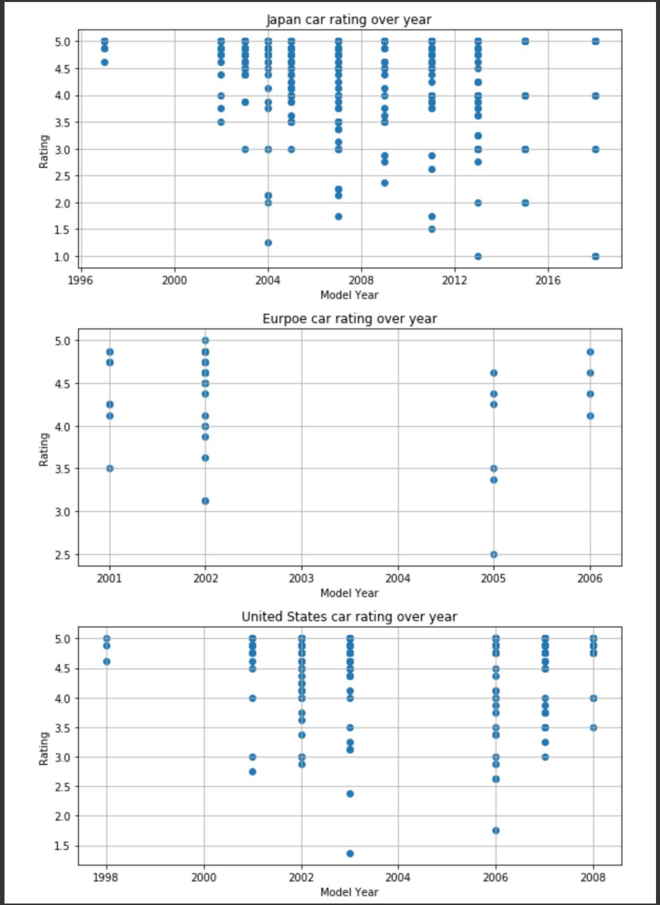


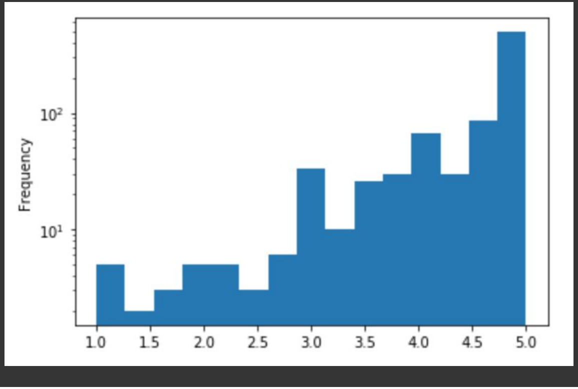
```
None
 99th %tile:  5.0
```

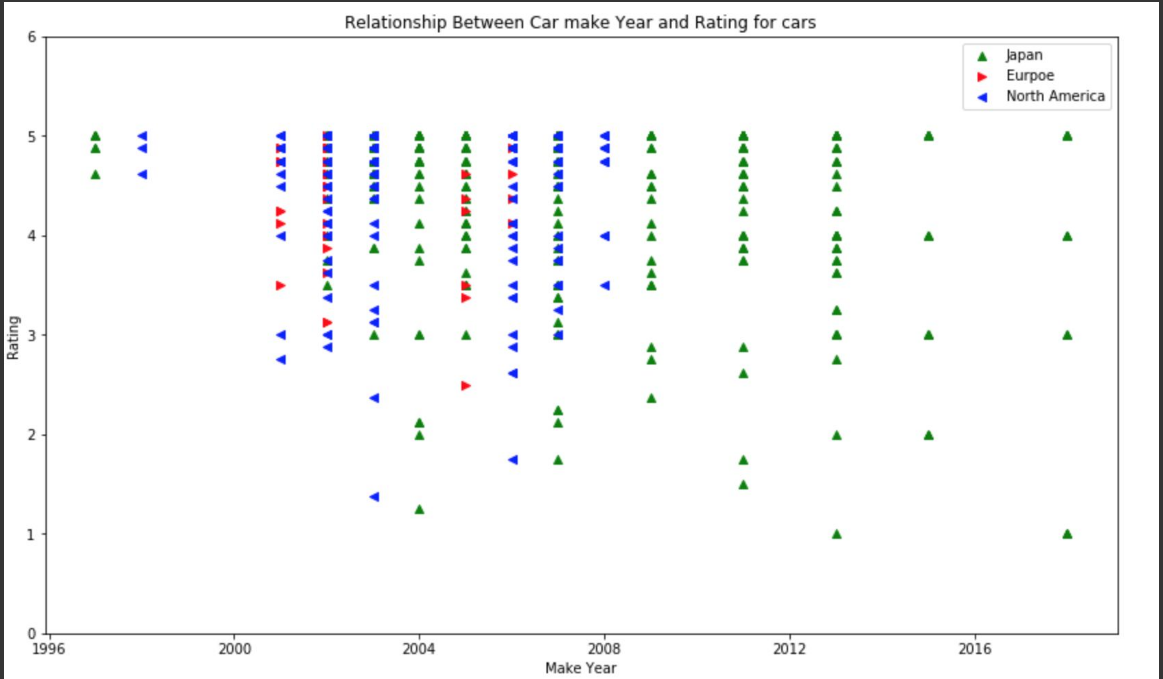


```
Data Types for Japense Car:  Unnamed: 0        object
Review_Date        object
Author_Name        object
Vehicle_Title      object
Review_Title       object
Review             object
Rating            float64
dtype: object
Data Types for European Car:  Unnamed: 0        object
Review_Date        object
Author_Name        object
Vehicle_Title      object
Review_Title       object
Review             object
Rating            float64
dtype: object
Data Types for American Car:  Unnamed: 0        object
Review_Date        object
Author_Name        object
Vehicle_Title      object
Review_Title       object
Review             object
Rating            float64
dtype: object
```

Some Graphs


Japan car rating over year
Rating
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1996
2000
2004
2008
2012
2016
Model Year
Eurpoe car rating over year
Rating
5.0
4.5
4.0
3.5
3.0
2.5
2001
2002
2003
2004
2005
2006
Model Year
United States car rating over year
Rating
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1998
2000
2002
2004
2006
2008
Model Year


Frequency
10^2
10^1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Relationship Between Car make Year and Rating for cars
Japan
Eurpoe
North America
Rating
6
5
4
3
2
1
0
1996
2000
2004
2008
2012
2016
Make Year

What can we conclude

Cars across region and over years tends to get consistent reviews.

Happy customers tends to review more than unhappy. 99% of reviews were rated as 5

Some inconsistencies in data shows that top reviewers only provided rating of 5 and they have all of their rating for specific car and in same year

Rating as integer are not best indicator, so keep it fractional value

The Rating is not evenly distributed (not a normal distribution). It's positively skewed

Big share of the information in the dataset is in form of descriptive set. Which makes it a good candidates for sentiment analysis.

Learning about reviewer's review pattern is not possible with the data provided as most of the review are provided anonymously

Key Observations

1. Data has many columns, big chunk of the information in the dataset is in form of descriptive set. Which makes it a good candidates for sentiment analysis.
2. There is enough data about 5% of total data where we have enough information to look into ratings and explore how rating were awarded, what things influenced the rating like Where car technology originated e.g. Asia (japan), Europe or USA.
3. The Rating dataset (subset of data cleaned up for Rating analysis) has some anolmolies
    a. The Rating is not evenly distributed (not a normal distribution). It's more negatively skewed
    b. Lot of review were provided anonymously, so making it difficult to identify reviewer pattern.
    c. Most of reviewers have reviewed same car for multiple times (1-5), so we cannot predict the reviewer bias about a given car or any comparison for same reviewer reviewing different cars. Which kind of make sense that user in japan would not have multiple cars to use and provide review.
    d. There were few reviewer anomalies where in one year few reviewers have reviewed 2000-3000 review for same car and all 5 rating.
    e. Rating density distribution increases from 4 - 5.
4. Different cars across region has consistent rating across the decade. Average rating remained same.
5. Japan cars have slightly higher 99th percentile (5.0) rating and European / American (4.5 - 4.8)

Next Steps

1) Extract additional data which can be used as features for analyzing and predicting the ratings, for example
   a) From Car Title extract
      i) Engine Power
      ii) Doors in Car
      iii) Gasoline (Petrol vs Diesel)
2) Include all cars datasets, and do more analysis to see if we can find reviewers are reviewing more than one car model or not.
3) Create a model which can predict a car rating based on car Make Year, Model, Engine Type and number of doors

Warning

If we want to include all 64 different car models and all data, it would need significant computing resources.

Feedback

Share your feedback to improve my analysis