Springboard Data Science Career Track
Capstone project 1 Milestone Report
By Qi Wu
Nov 24 2019

1. Write a capstone project 1 milestone report (Google Doc, 5-6 pages) and include the following:
    1. Problem statement: Why it's a useful question to answer and for whom (source this from your proposal)

The problem I want to solve is Amazon instant reviews classification. My clients are marketing companies, and they care about this because through this model, we can potentially study connections between reviews and what people care and do not care about.

    2. Description of the dataset, how you obtained, cleaned, and wrangled it (source this from your data wrangling report)

The dataset I got is from Amazon. They have a review database for instant videos from 2000-2014 with 37126 data points and 9 features. The feature names are:

reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B

asin - ID of the product, e.g. B000H00VBQ

reviewerName - name of the reviewer

helpful - helpfulness rating of the review, e.g. 2/3

reviewText - text of the review

overall - rating of the product

summary - summary of the review

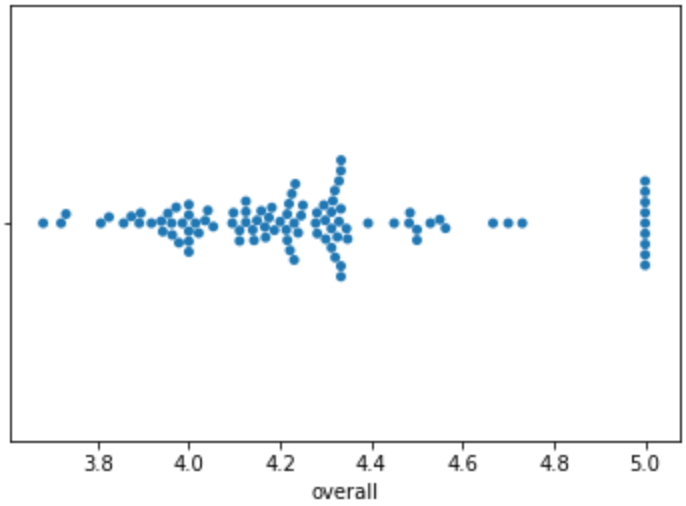
unixReviewTime - time of the review (Unix time)

reviewTime - time of the review (raw)

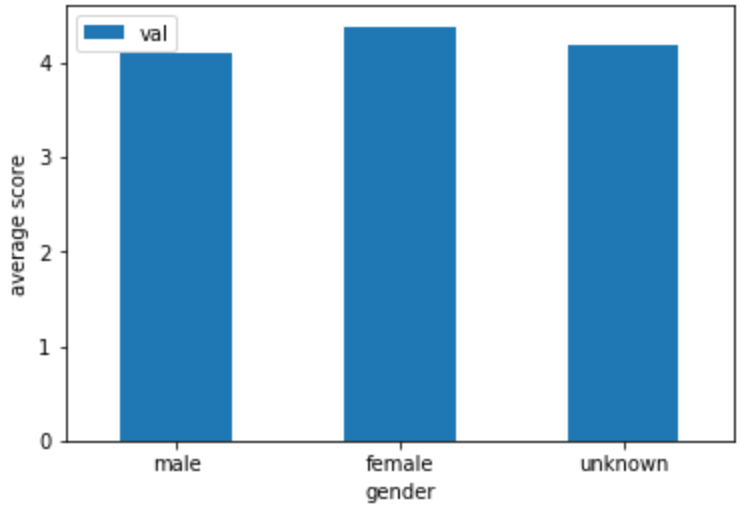
So at the data wrangling part, I imported the dataset to Pandas dataframe, cleaned the Null values rows, insert the time index and gender feature for further analysis.

3. Initial findings from exploratory analysis (source this from your data story and inferential statistics reports)
    1. Summary of your findings
    2. Visuals and statistics to support those findings

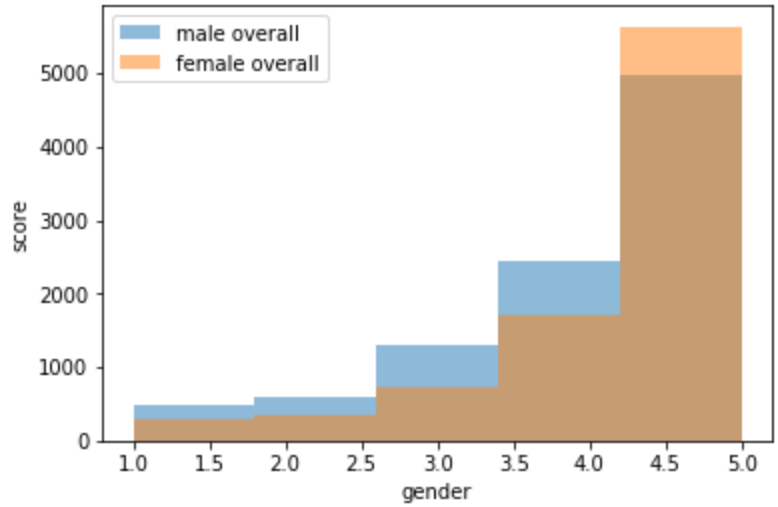
90% of the reviews are about reviews with a score of 3 and all of the monthly average score is above 3



I used the name of the reviewer to infer the gender using Gender_guesser package. The average score of male and female are both around 4.



There is a difference between them.



Comparing the average between male (4.2) and female (4.4) reviewers we find that there is a difference of (0.2). I studied whether this difference is statistically significant, as follows.

Null Hypothesis: (The average overall review score from women and men has no significant difference)

Alternative Hypothesis: (The average overall review score from women and men has significant difference)

Threshold for acceptance: 0.01

I then used the F distribution, and found that the critical value is (290), and the p-value is (1.3e-64). Since the p-value is less than the threshold, then I can reject the null hypothesis, and therefore the difference observed is statistically significant.

There is no linear regression relation between time and average review score.
1) Based on the value of R-squared, it showed the relationship between the time and the average review score is weak. Only around 25% of the average score can be explained by the time factor.
(2) the relationship between the time and average score is not significant when the threshold for acceptance is 0.01, which is far more big than the P value of F-statistics 9.34e-08.

OLS Regression Results

```
==============================================================================
Dep. Variable:                overall   R-squared:                       0.253
Model:                            OLS   Adj. R-squared:                  0.246
Method:                 Least Squares   F-statistic:                     33.27
Date:                Mon, 25 Nov 2019   Prob (F-statistic):           9.34e-08
Time:                        19:02:01   Log-Likelihood:                -10.455
No. Observations:                 100   AIC:                             24.91
Df Residuals:                      98   BIC:                             30.12
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.8312      0.103     46.894      0.000       4.627       5.036
Datenum       -0.0002   2.87e-05     -5.768      0.000      -0.000      -0.000
==============================================================================
Omnibus:                        1.528   Durbin-Watson:                   1.081
Prob(Omnibus):                  0.466   Jarque-Bera (JB):                1.349
Skew:                           0.133   Prob(JB):                        0.510
Kurtosis:                       2.498   Cond. No.                     1.36e+04
==============================================================================
```

Next Steps:

1. Build a Multinomial Naive Bayes model to show the relationship between the review polarity and its text.
2. Use another machine learning algorithms like Logistic Regression to build a model too and compare which model is better.