Capstone Project 1

Consolidated Report

Rate A Read With goodreads

goodreads

Introduction:

Goodreads is a social cataloging website for people who love Books.Users can just sign up and then create a reading list or update the books they have read or

currently reading or even write a review. They can also form their own groups of book suggestions, surveys, polls, blogs, and discussions.

Proposal/Problem Statement:

In this project, I will explore Science Fiction/Fantasy Genre. I will collect Books and Author Details from *goodreads* and will analyze different features to determine what makes a book popular or what are the determinants in a book which earns a good rating and finally I will **predict** the Average Rating of a Book.

Apart from the regular and important concepts of Data Wrangling, Exploratory Analysis and fitting an ML algorithm ,I will try some other interesting concepts like below:

1. Collect Data using an API
2. Database Design
3. ETL (Extract Transform Load)

Data Collection:

Books and the corresponding authors details are collected from goodreads using an API. But to use the GoodReads API, we need to register for a developer key. The key can be registered on https://www.goodreads.com/api/keys.

Books are searched using the **search_books** method of the API passing Genre as a parameter. In this project, extracted books details for the following genres:
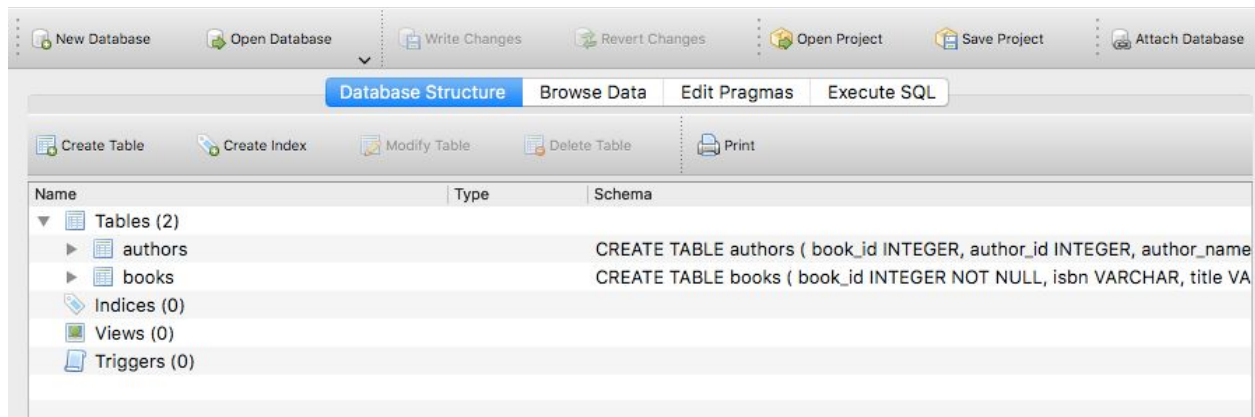
1. Science fiction
2. Science-fiction-fantasy
3. Science-fiction-romance
4. Apocalyptic
5. Space
6. Dystopia
7. Aliens
8. Fantasy

Database Design:

In this project, used a SQLite3 Database to load the data.

Extracted both the Books details and the Author information from goodreads for a particular genre.

Created 2 tables as below:

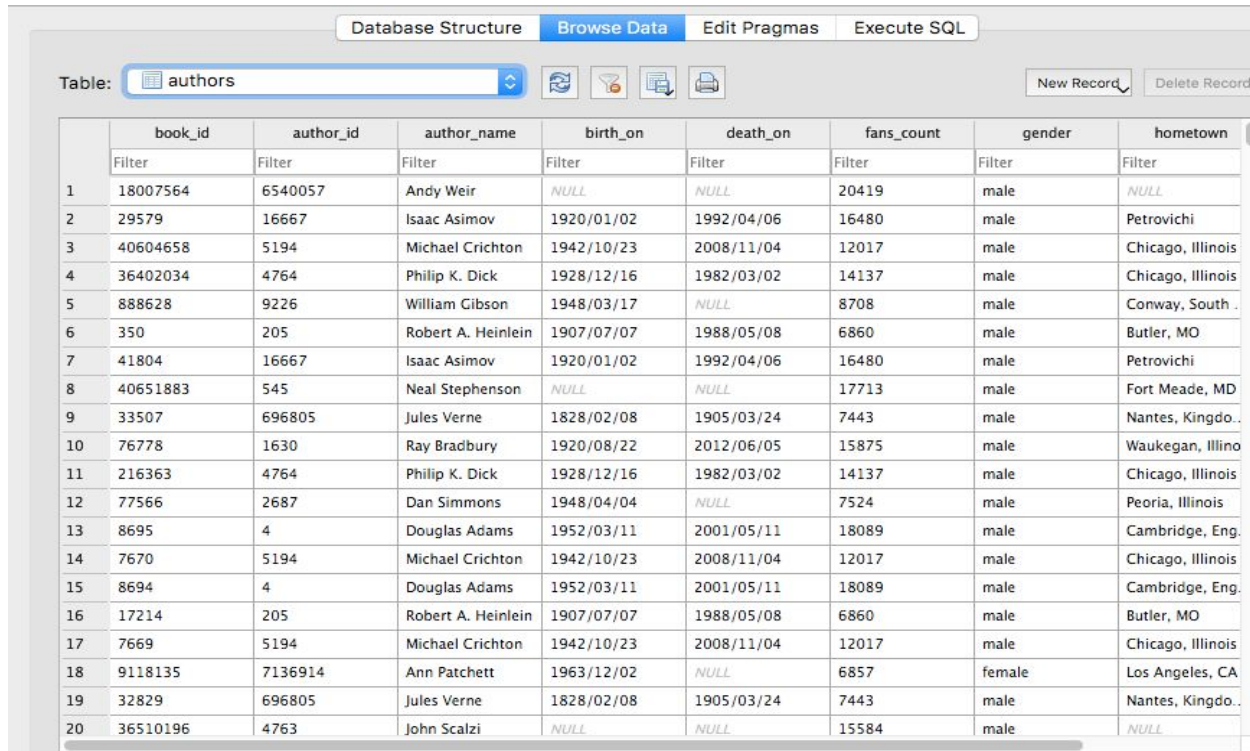1. **Books** - To store the book details where book_id is the Primary key

2. **Authors** - To store the author details where book_id from Books table is the Foreign key

	book_id	author_id	author_name	birth_on	death_on	fans_count	gender	hometown
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	18007564	6540057	Andy Weir	NULL	NULL	20419	male	NULL
2	29579	16667	Isaac Asimov	1920/01/02	1992/04/06	16480	male	Petrovichi
3	40604658	5194	Michael Crichton	1942/10/23	2008/11/04	12017	male	Chicago, Illinois
4	36402034	4764	Philip K. Dick	1928/12/16	1982/03/02	14137	male	Chicago, Illinois
5	888628	9226	William Gibson	1948/03/17	NULL	8708	male	Conway, South .
6	350	205	Robert A. Heinlein	1907/07/07	1988/05/08	6860	male	Butler, MO
7	41804	16667	Isaac Asimov	1920/01/02	1992/04/06	16480	male	Petrovichi
8	40651883	545	Neal Stephenson	NULL	NULL	17713	male	Fort Meade, MD
9	33507	696805	Jules Verne	1828/02/08	1905/03/24	7443	male	Nantes, Kingdo..
10	76778	1630	Ray Bradbury	1920/08/22	2012/06/05	15875	male	Waukegan, Illino
11	216363	4764	Philip K. Dick	1928/12/16	1982/03/02	14137	male	Chicago, Illinois
12	77566	2687	Dan Simmons	1948/04/04	NULL	7524	male	Peoria, Illinois
13	8695	4	Douglas Adams	1952/03/11	2001/05/11	18089	male	Cambridge, Eng.
14	7670	5194	Michael Crichton	1942/10/23	2008/11/04	12017	male	Chicago, Illinois
15	8694	4	Douglas Adams	1952/03/11	2001/05/11	18089	male	Cambridge, Eng.
16	17214	205	Robert A. Heinlein	1907/07/07	1988/05/08	6860	male	Butler, MO
17	7669	5194	Michael Crichton	1942/10/23	2008/11/04	12017	male	Chicago, Illinois
18	9118135	7136914	Ann Patchett	1963/12/02	NULL	6857	female	Los Angeles, CA
19	32829	696805	Jules Verne	1828/02/08	1905/03/24	7443	male	Nantes, Kingdo..
20	36510196	4763	John Scalzi	NULL	NULL	15584	male	NULL

Extract-Transform-Load (ETL):

Extract: Extracted Data from a homogenous source i.e. goodreads

Transform: Transformed Data into a proper storage format/structure

Load: Inserted Data into the Target Database Tables(Books,Authors)

Data Wrangling:

The Data Loaded from Database into a Dataframe and below are the features:

```
df_details.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 9576 entries, 0 to 9575
Data columns (total 19 columns):
book_id             9576 non-null int64
isbn                7864 non-null object
title               9576 non-null object
total_pages         8848 non-null float64
average_rating      9576 non-null float64
ratings_count       9576 non-null int64
reviews_count       9576 non-null int64
publication_date    8029 non-null object
publisher           8521 non-null object
popular_shelves     8029 non-null object
book_description    9370 non-null object
author_id           9576 non-null int64
author_name         9576 non-null object
birth_on            3482 non-null object
death_on            1201 non-null object
fans_count          9576 non-null int64
gender              8321 non-null object
hometown            5729 non-null object
works_count         9576 non-null int64
dtypes: float64(2), int64(6), object(11)
memory usage: 1.5+ MB
```

Data Cleaning

- ❖ Convert the Gender column to category
- ❖ Convert the Date columns to Dates
- ❖ Handling Missing Data
  - For **total_pages** column, the missing values are filled with the MEAN of the total pages of the other records.
  - For **fans_count** column, the missing values are filled with the MEAN of the fans count of the other records.
  - For **popular shelves**, the mission value is filled with "No_Tags".
  - **Gender** Missing Values are filled with the Forward fill method.
  - Missing **Book_description** column is filled with a constant value "No_Description".

Feature Extraction

Tags:

Popular_shelves column of the books are used to fetch tags of each book by using the steps below:

- Join shelves of each record to get all the shelves.
- Exclude not so important Tags
- Fetched the Most Common Tags Value
- Create new Tags like "classics", "thriller", "romance" ,"paranormal" , "humour", "dystopian", "historical", "comics" and put True/False for each record

Bag of Words from Book Description:

Bag Of Words is used to extract features from Book_Description column using the steps below:

1. The records are converted to lowercase
2. HTML tags are removed from the records
3. Punctuations are removed from the records
4. Trailing spaces are removed from the records
5. Spaces in between words are removed from the records
6. Numbers are removed from the records
7. English stop words are removed
8. Tokenization, Stemming and Lemmatization process are used to clean the data
9. CountVectorizer method is used to get counts of each word

Exploratory Data Analysis:

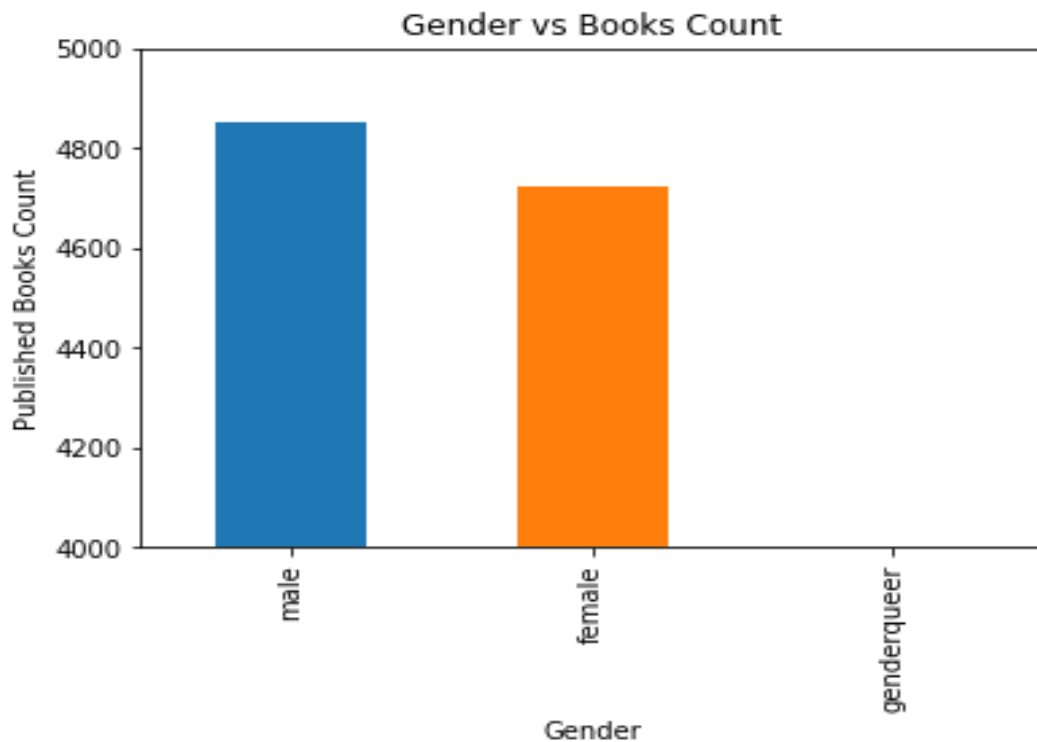In this project, we are predicting the average rating of a book in Science Fiction Genre.

Average Rating:
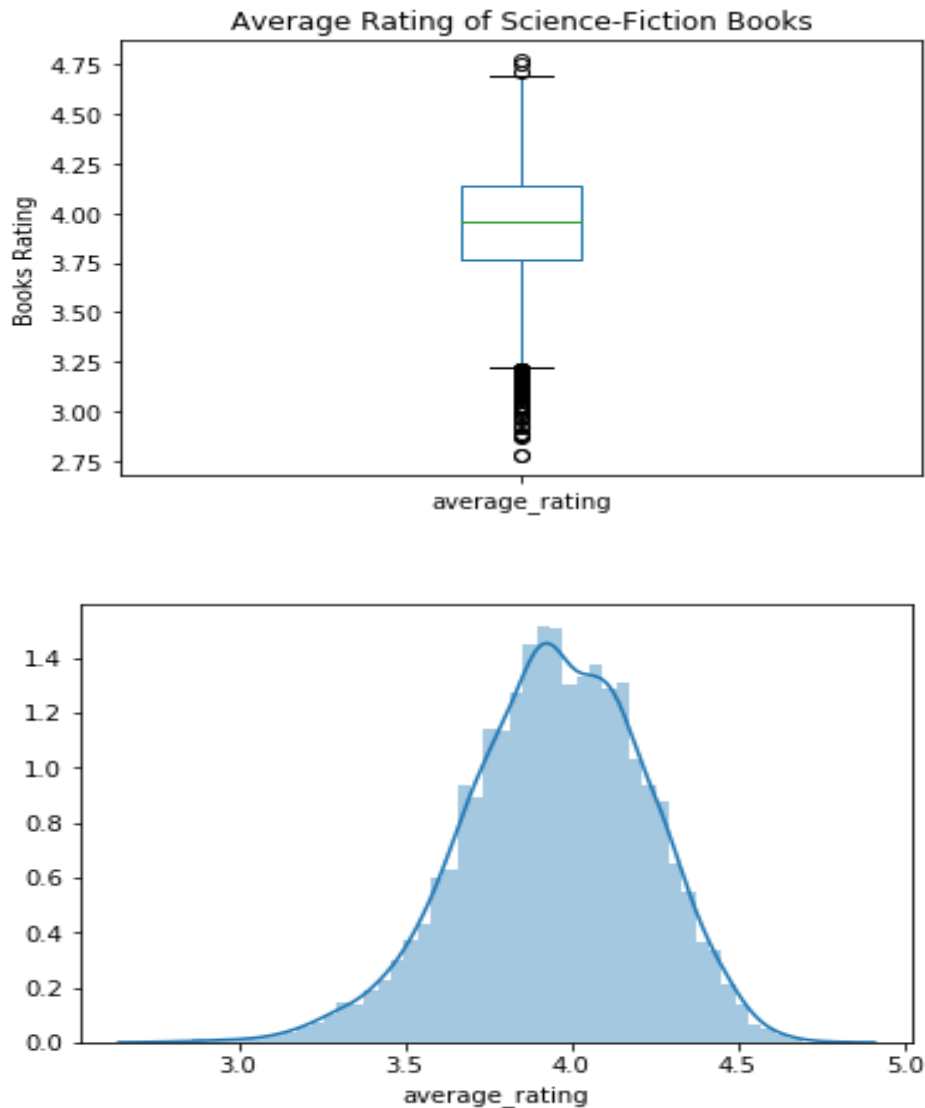
The rating of the book varies from 2.75 to 4.75 with a mean value around 4 and is Normally Distributed.

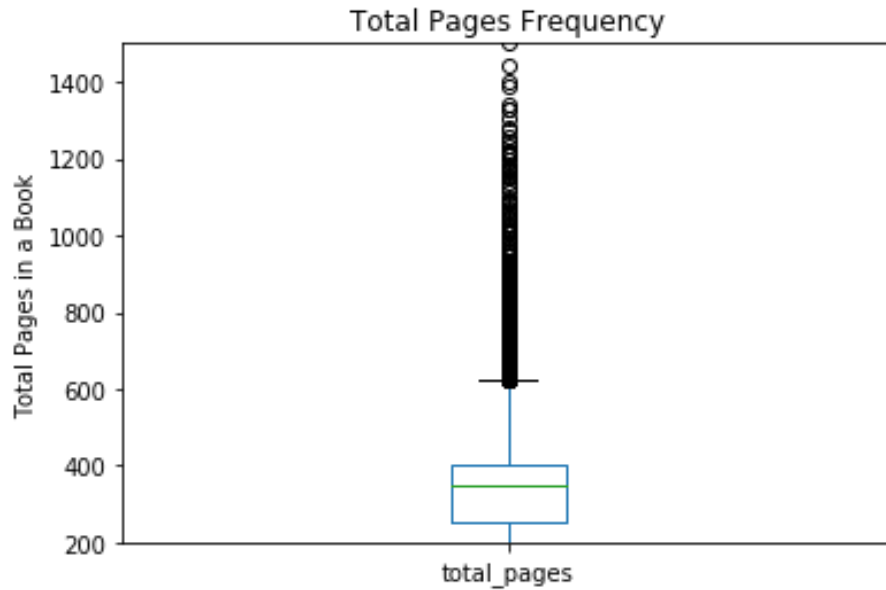**Gender**:

In this project, the Gender of the Author is playing an important role and it seems that there are more Male author than Female authors in the world of Science Fiction.
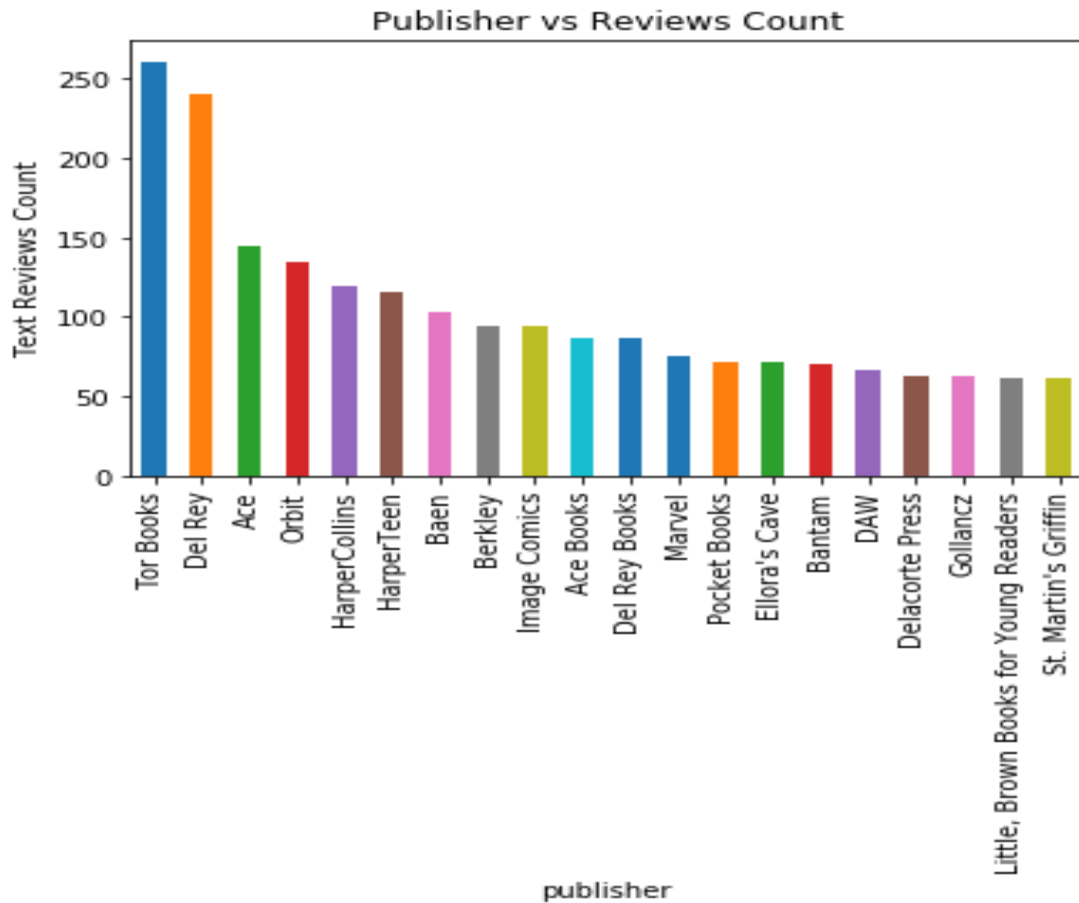
Total Pages:

Total Pages Frequency

In Science Fiction, there are a couple of books which consist of many pages. But, for most of the books, the page count is at a mean of around 350.

Publishers:

Publisher vs Reviews Count

"Tor Books" is the most popular publisher in the world of Science Fiction.

Sub Genre:
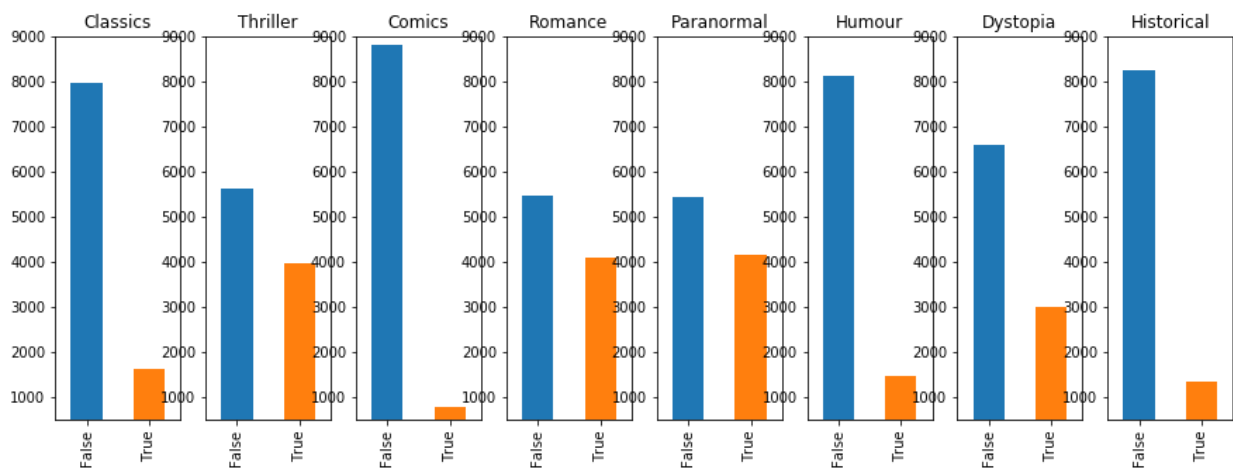
Science Fiction is a broad Genre. Under Science Fiction, there are some sub-categories and the above plot shows the distribution.

City:

City Vs Science-Fiction Books Count

Can a City influence a creation? The above plot shows that it can! Chicago and Cleveland have given birth to most of the Science Fiction Creations.

Authors:

Authors who created most of the "Classics" in Science Fiction is "Robert A. Heinlein".

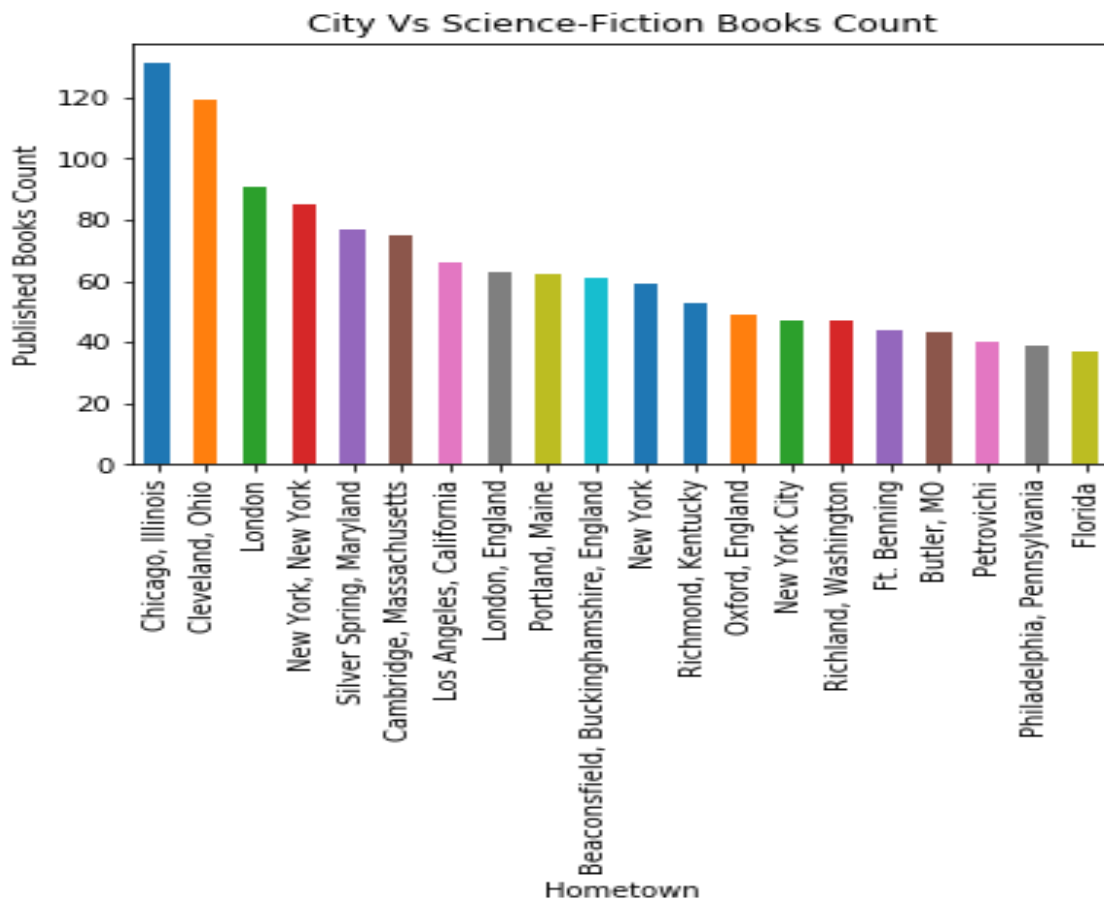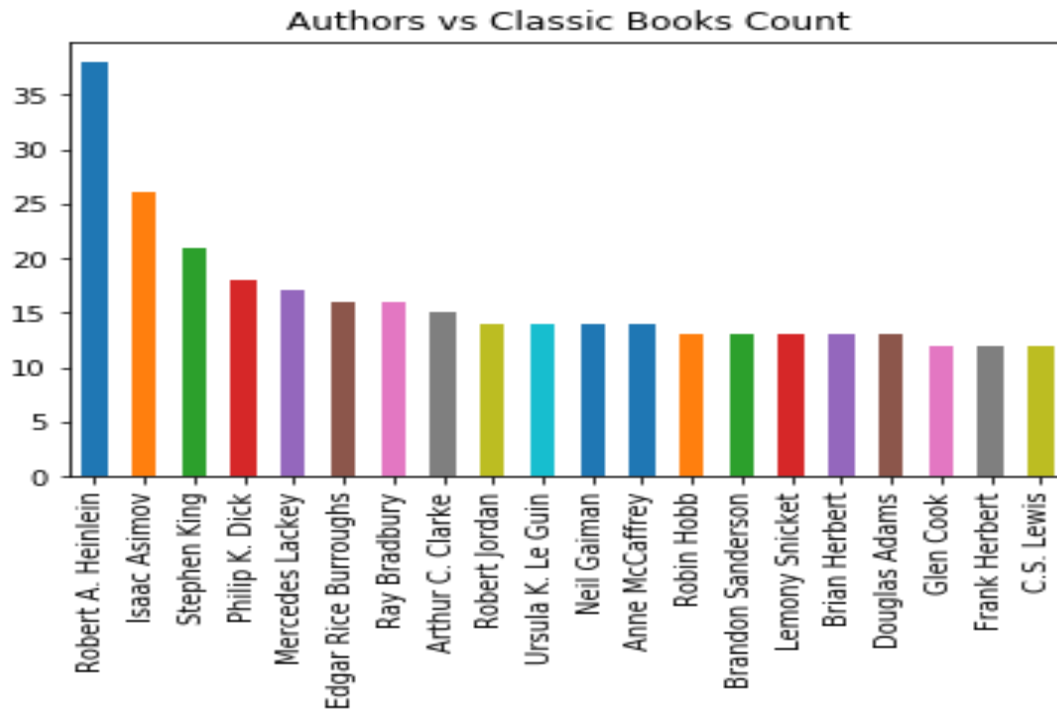**Best Reads in Science Fiction:**

title	average_rating	author_name	publisher
Weirdos from Another Planet! (Calvin and Hobbes #4)	4.71	Bill Watterson	Andrews McMeel Publishing
Harry Potter Series Box Set (Harry Potter, #1-7)	4.75	J.K. Rowling	Arthur A. Levine Books
Black Dagger Brotherhood: Boxed Set #1-6	4.69	J.R. Ward	null
Words of Radiance (The Stormlight Archive, #2)	4.77	Brandon Sanderson	Tor Books
A Court of Mist and Fury (A Cou of Thorns and Roses, #2)	4.68	Sarah J. Maas	Bloomsbury USA Childrens
Saga: Book One	4.68	Brian K. Vaughan	Image Comics

The top 5 Best Reads in Science Fiction are displayed in the above chart.

Publication Year:
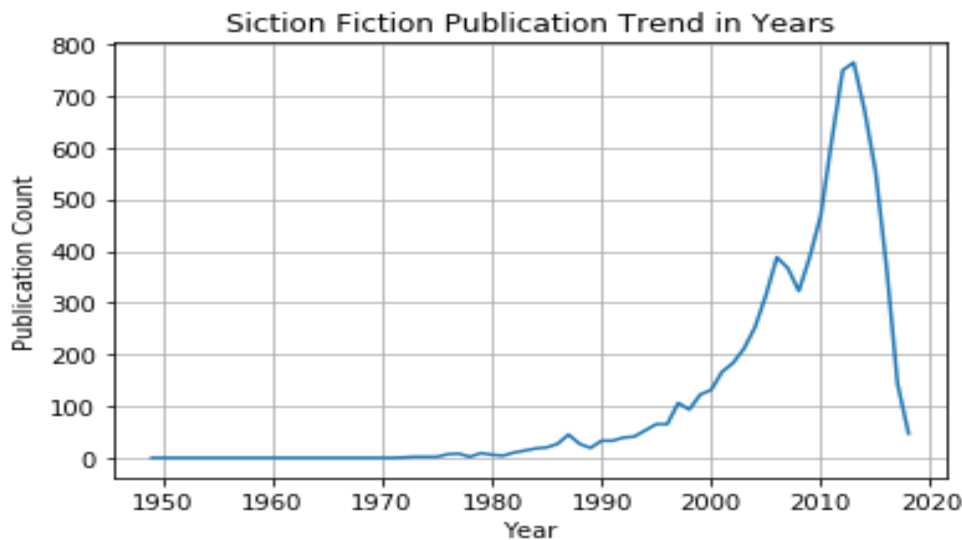
The above plot shows the count of Science Fiction Publication as per year. It seems that this particular genre started gaining popularity in the 21st century.

Machine Learning Algorithms:

Train a Linear Regression With statsmodels

Statsmodel is a Python library designed for more statistically-oriented approaches to data analysis. It has some built in support for many of the statistical tests to check the quality of the fit and a dedicated set of plotting functions to visualize and diagnose the fit.

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.441
Model:                            OLS   Adj. R-squared:                  0.272
Method:                 Least Squares   F-statistic:                     2.612
Date:                Fri, 14 Jun 2019   Prob (F-statistic):           5.05e-200
Time:                        10:13:39   Log-Likelihood:                 1641.5
No. Observations:                9576   AIC:                             1155.
Df Residuals:                    7357   BIC:                         1.706e+04
Df Model:                        2218
Covariance Type:            nonrobust
------------------------------------------------------------------------------
```

The R-Squared Value is **0.441** and the Adjusted R- Squared value is **0.27**.

Fitted Values vs Residuals
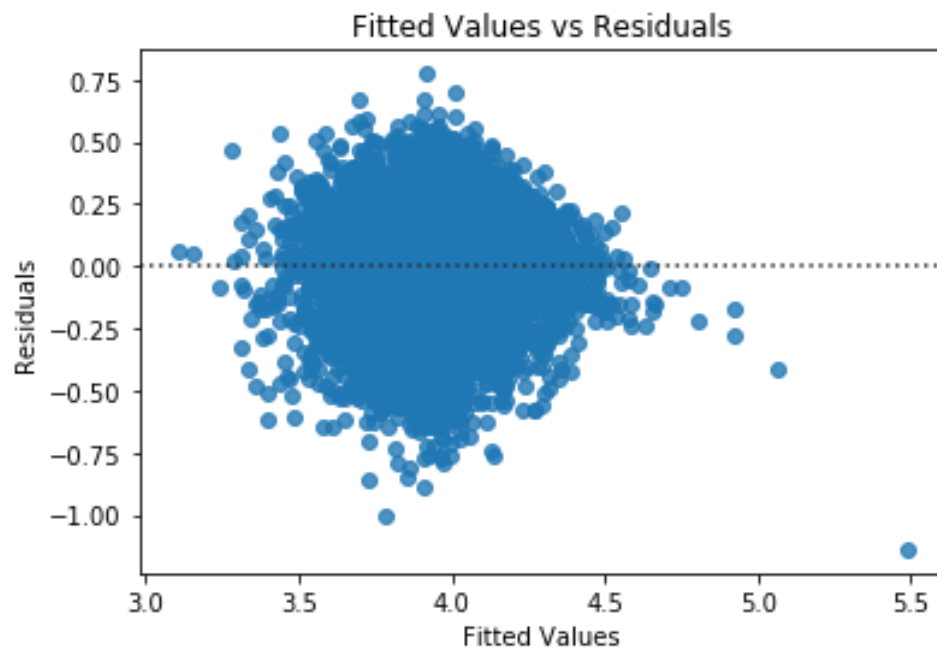
Train a Linear Regression Model With scikit-learn

Scikit-learn is a powerful Python module for machine learning. In this project I have explored the sklearn.linear_model module which contains *"methods intended for regression in which the target value is expected to be a linear combination of the input variables"*.

Actual Values Vs Predicted Values For Linear Regression
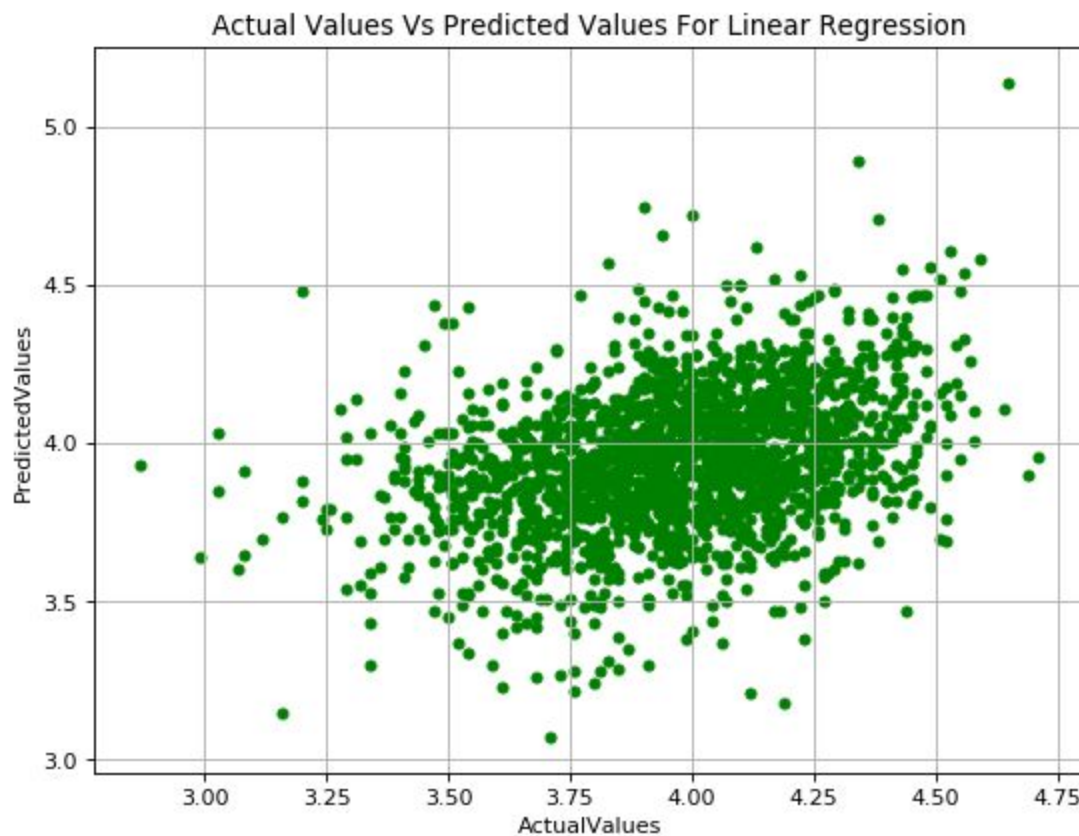
The Linear Regression Model did not perform that well and R2 score is -0.1. So, the next step is to look for some better Fit.

Train a Decision Tree Regression Model

Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility.

The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]
2. Result [End Nodes]

Actual Values Vs Predicted Values For Decision Tree Regression
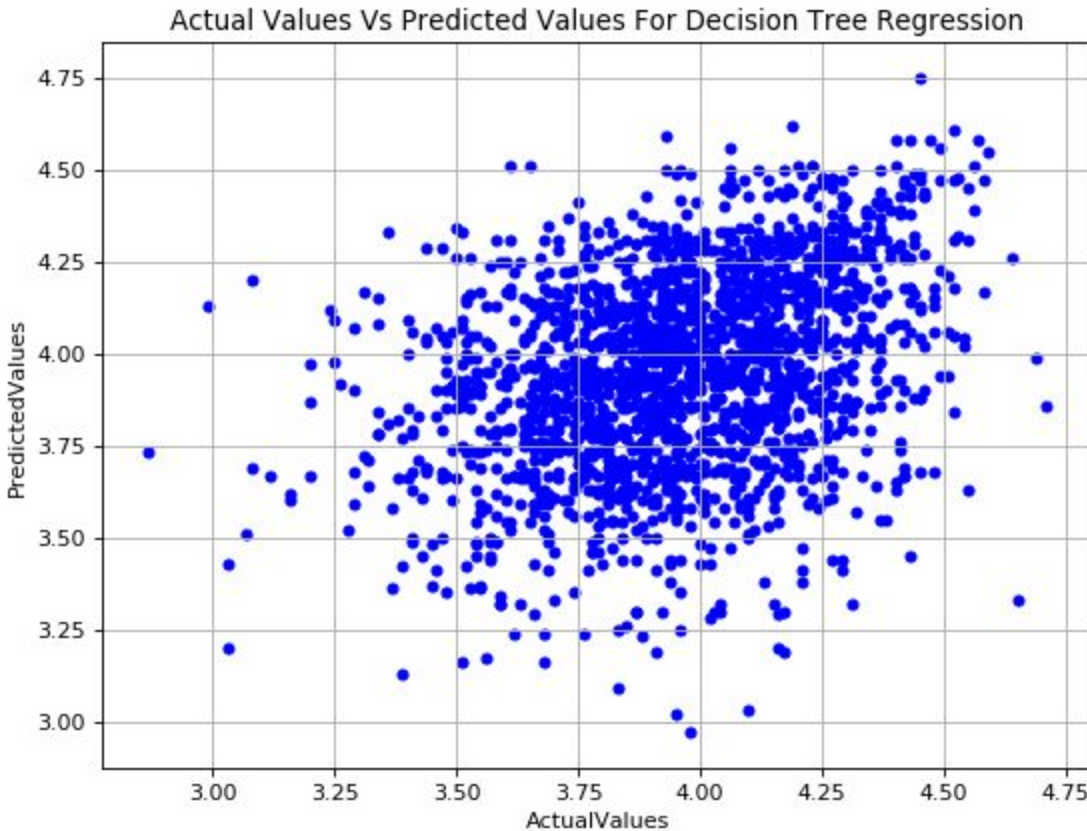
Again this model did not perform that well in our current Dataset and resulted in R2 score of **-0.300.**

Train a Random Forest Model

A Random Forest combines multiple decision trees in determining the final output rather than relying on individual decision trees. The Model Outputs the mean prediction (regression) of the individual trees.

Random Search Cross Validation
Used the below grid with 3 fold Cross Validation for Random Search:

{'bootstrap': [True, False],
 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]}
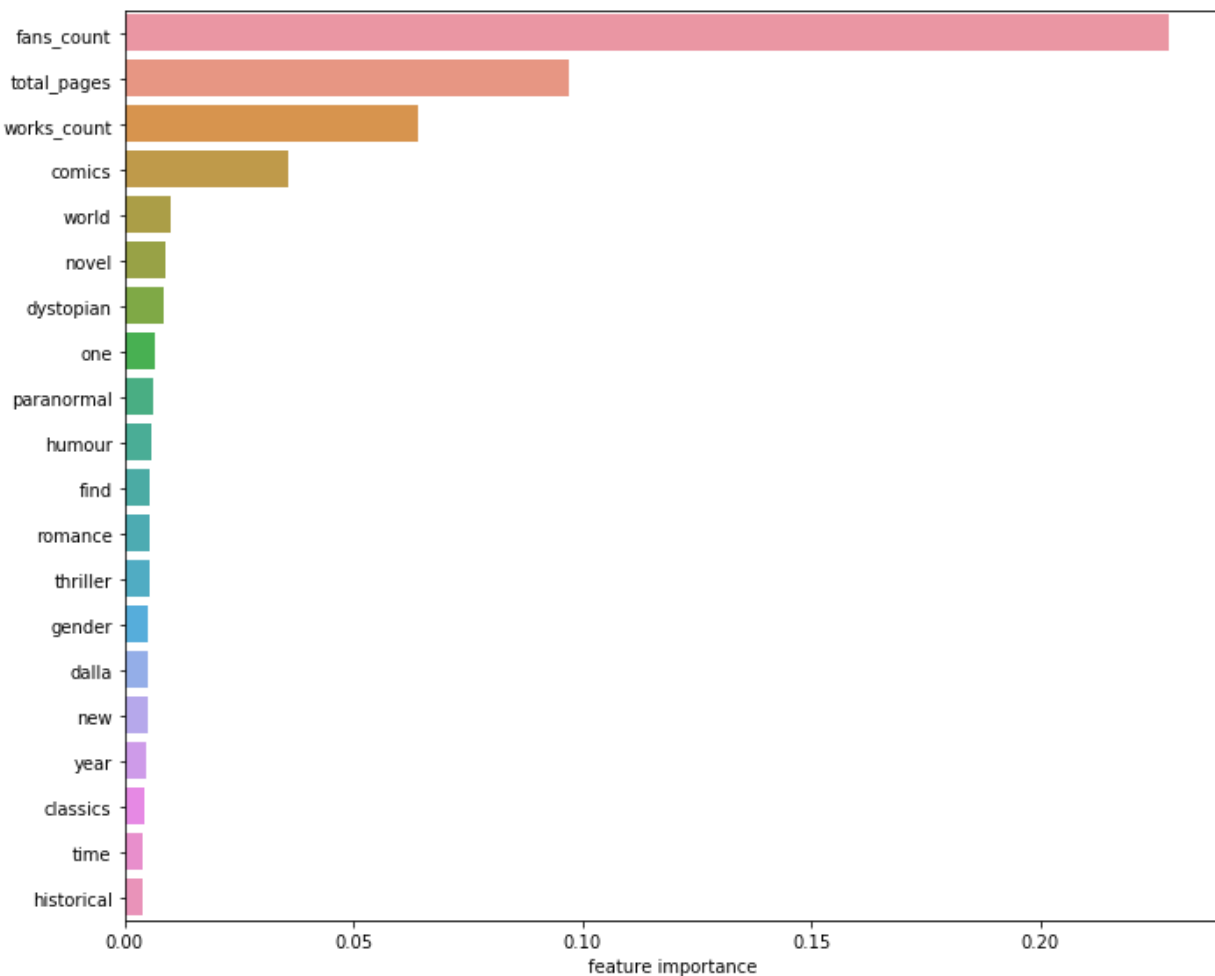
Best Parameters set found:
{'n_estimators': 900, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': None, 'bootstrap': True}

Fitted a Random Forest Model with the best parameters based on the results of Random Search.

Feature Importance as explored by RF Model

Reduced Features to explore only features with 95% importance and below are the top variables.

Fitted a Final Random Forest Model using Best Parameters provided by Grid Search and the Reduced Features. This Model did reasonably well and the R2 score is 0.34.

Train a Support Vector Regression Model

Support Vector Regression (SVR) is a regression algorithm, and it applies a similar technique of Support Vector Machines (SVM) for regression analysis. As we know, regression data contains continuous real numbers. To fit such type of data, the SVR model approximates the best values with a given margin called ε-tube (epsilon-tube, ε identifies a tube width) with considering the model complexity and error rate.
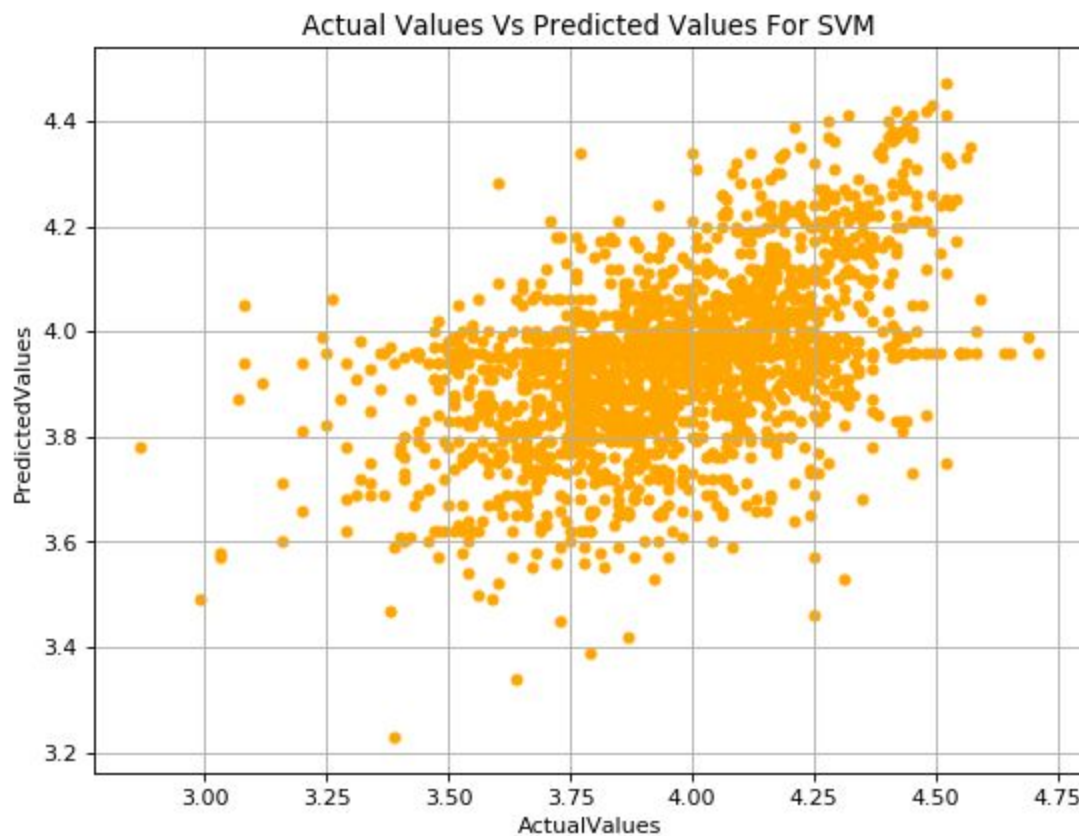
Compare Models

A well-fitting regression model results in predicted values close to the observed data values. The below Metrics are used for evaluation:

1. **R2_score**: This function computes the coefficient of determination, usually denoted as R². It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model. The value usually varies between 0 (worst fit) and 1 (best fit) and even it can be a negative value(because the model can be arbitrarily worse).
2. **Mean Absolute Error (MAE)** is the mean of the absolute value of the errors.
3. **Mean Squared Error (MSE)** is the mean of the squared errors. The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.
4. **Root Mean Squared Error (RMSE)** is the square root of the mean of the squared errors.

Model	R2 Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
OLS Stats Model	0.272			
Sklearn Linear Regression Model	-0.100	0.218	0.079	0.28
Decision Tree	-0.300	0.235	0.0932	0.30
Random Forest Model	0.343	0.169	0.047	0.217
Support Vector ReRegressor	0.240	0.179	0.054	0.233

In this project, I have implemented different models to predict the Average_Rating of a book and the *Random Forest Model did the best*.

Conclusion:
There is a popular saying:

"It is far better to foresee even without certainty than not to foresee at all."

Machine Learning Algorithms have evolved with time and made it possible to foresee future with more certainty. The Machine Learning approach involves learning from DATA by identifying patterns and thus using them to automatically make some predictions.

In this project, I have used Multivariate Regression Algorithms to predict an Average Rating of a Book. Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables(the predictors).

It is difficult to predict the exact rating of a book since a Book gains popularity over time and the Rating of a book gets better with time and with more readers. Keeping those constraints in mind, in this project, I have tried to explore different features which a Writer can consider checking out before launching a book on Scienec Fiction/Fantasy.

Thank You!