Lineaje Data Analytics Challenge

MongoDB Sample Datasets

presented by
Jacob Farr

I. Introduction

Corey Walker of Lineaje extended me a challenge as part of the Junior Data Analyst interview process. That challenge came with specific expectations and opportunities. The opportunity was to dig deeper into MongoDB and Python/R programming languages using the sample data listed here: https://www.mongodb.com/docs/atlas/sample-data/#std-label-load-sample-data. The expectations were to create a free mongodb account, load the sample dataset, generate interesting correlations from the data, and to explain my point of view on those observations. Corey expressed special interest in my thought process and ideally a final solution. Therefore, I wrote this executive summary to display both.

II. Load Sample Dataset

Data Preprocessing

There were a few steps which needed to be taken before the data could be worked with using the Python data analysis libraries Pandas, Seaborn, Numpy, and Matplotlib. First was to find a way to export some of the sample data from the MongoDB cluster into a .csv file using an app called MongoDBCompass. Once several of the databases were downloaded onto a local computer, some cleanup was completed. Duplicates were dropped and NaN values were handled. At this point, data exploration and analysis seemed appropriate.

III. Data Exploration

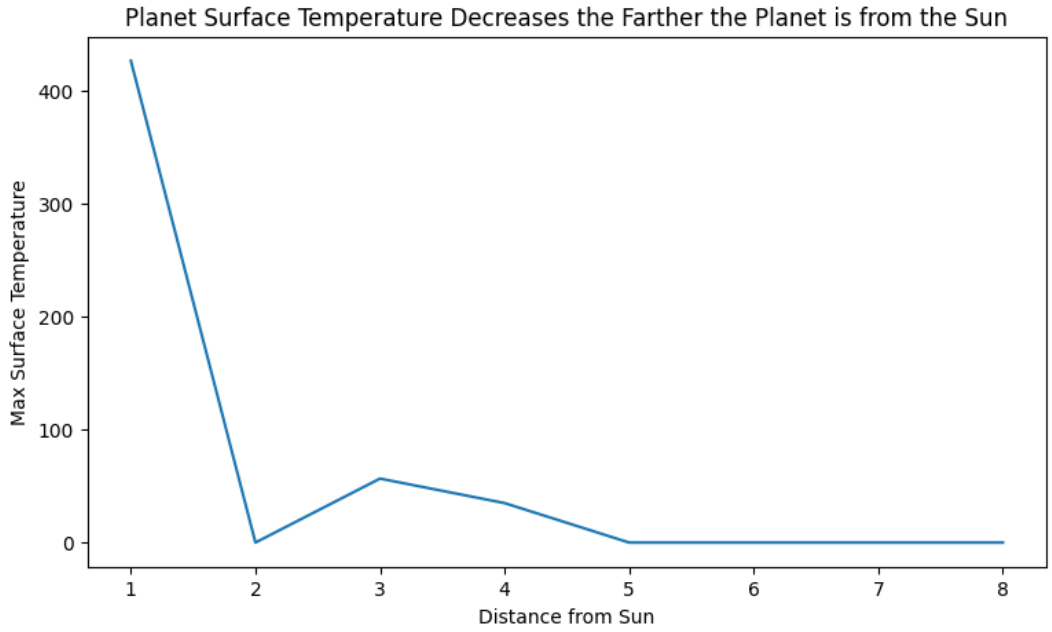
Methods

The Python open-source libraries Pandas and Seaborn were heavily relied upon through the course of this analysis. Noteworthy methods include .corr(), which computes pairwise correlation of columns, excluding NA/null values. This made it very easy to find positive and negative correlations between columns within a dataset. Here is an example of .corr():

	hasRings	orderFromSun	surfaceTemperatureC.max	surfaceTemperatureC.mean	surfaceTemperatureC.min
hasRings	1.000000	0.872872	-0.468630	-0.707493	0.736739
orderFromSun	0.872872	1.000000	-0.629728	-0.758826	0.643396
surfaceTemperatureC.max	-0.468630	-0.629728	1.000000	0.176518	-0.757764
surfaceTemperatureC.mean	-0.707493	-0.758826	0.176518	1.000000	-0.123599
surfaceTemperatureC.min	0.736739	0.643396	-0.757764	-0.123599	1.000000

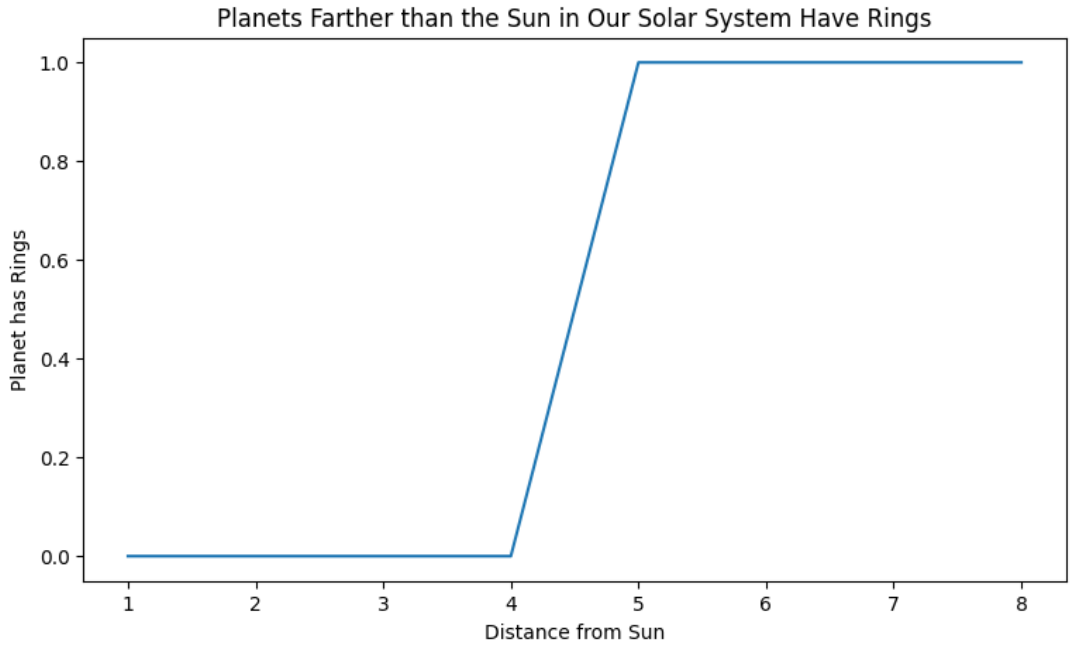
The closer a number is to 1, the more positive the correlation. The closer a number is to 0, the more negative the correlation. Once a high correlation was found in a dataset, then a visualization was created to represent that correlation in a chart. Following below are a few of the correlations discovered during data exploration. Please note that this is not a comprehensive analysis and herein I am simply providing, as requested, a few correlations from the sample data provided by mongoDB.

What is the surface temperature like on the planets in our solar system?



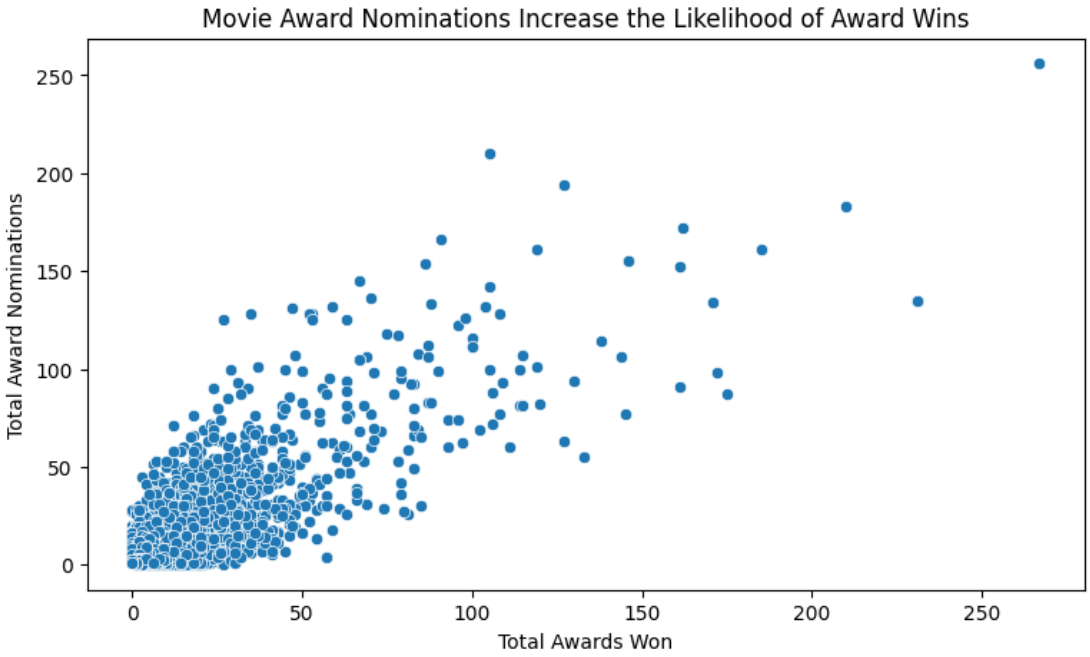
As you can see from the chart, planet surface temperature generally decreases the further a planet is from the sun. Between planets 1, 3, and 4, the temperature goes down with distance. However, please note that when I handled NaN values, I replaced the NaN values with a 0 throughout the entire dataset. Planets 2, 5, 6, 7, and 8 are possibly valued as 0 because their max surface temperature was not initially provided.

What determines whether or not a planet in our solar system has rings?

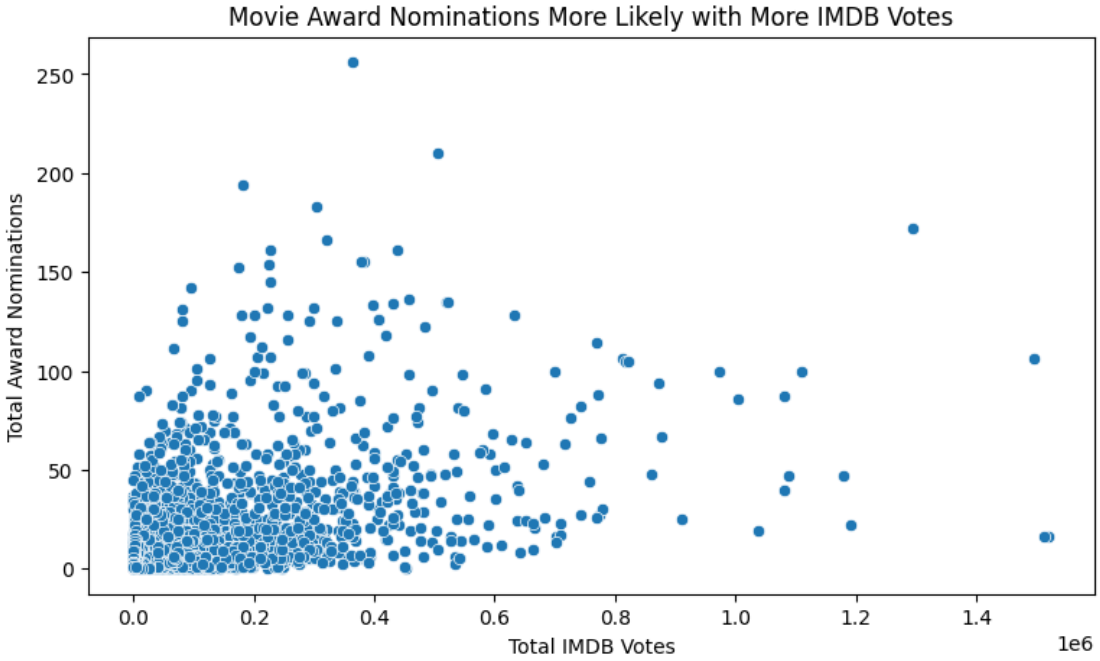


It might appear that planets farther from the sun tend to have rings. Why is this the case? Maybe planets get lonely so far out in space and need the rings to keep them company. Maybe distance has something to do with planetary rings. Your guess is as good as mine.

How can you increase the likelihood that your movie will win an award?

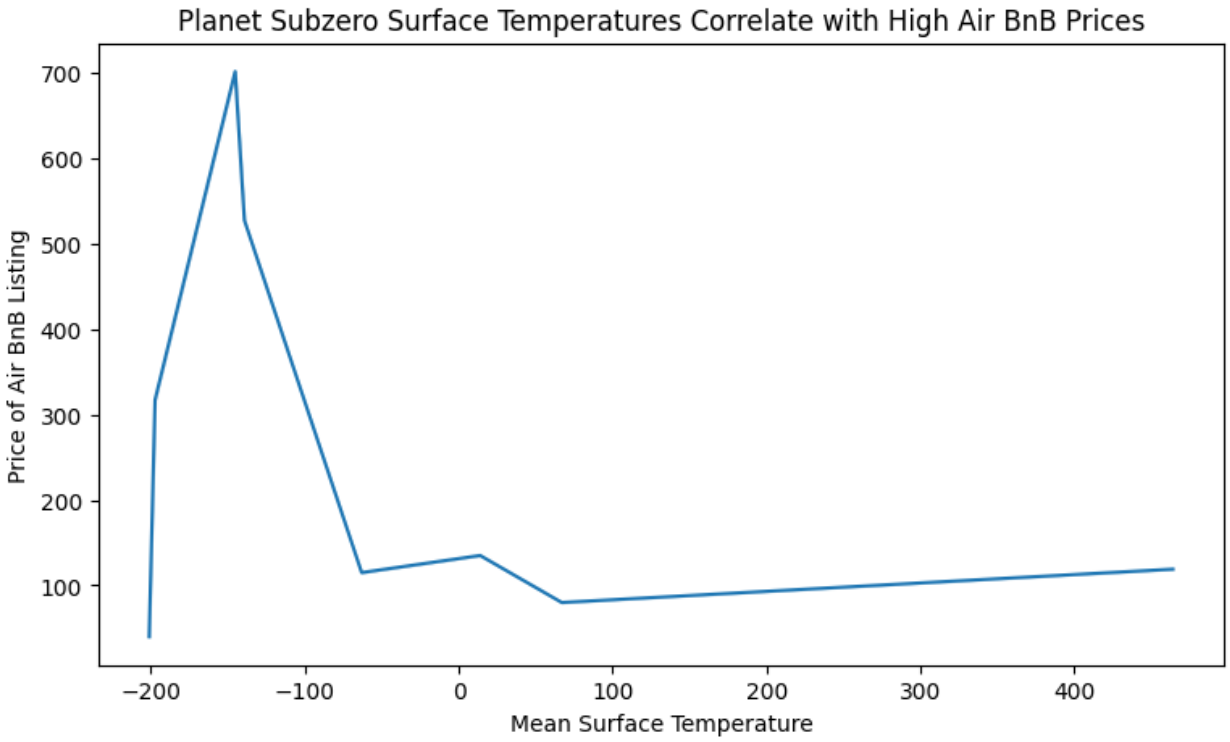


The number of award nominations is positively correlated with the number of awards won. This means that a movie that has been nominated for an award many times is more likely to win an award. Therefore, a movie director has nothing to lose by award nominations.



We also see that the IMDB Votes are loosely, but positively, correlated with the number of award nominations. That means a movie is more likely to be nominated with more votes on IMDB, regardless of whether they are good or bad reviews. If a director is having a bad time being nominated for an award then a milestone goal may be to obtain more IMDB votes.

Just for Fun: Have you ever wondered if planet surface temperatures correlates with the price of an Air BnB? Well today is your lucky day!



We can see from the chart that subzero surface temperatures are correlated with higher Air BnB prices. Who knew! You should probably check the surface temperature of Pluto before booking your next Air BnB. You might end up paying more than you bargained for!

IV. Python Notebooks and Citation Links

https://github.com/Morthais/_LineageChallenge

https://pandas.pydata.org/pandas-docs/stable/index.html

https://seaborn.pydata.org/

https://www.mongodb.com/developer/languages/python/pymongoarrow-and-data-analysis/

https://www.mongodb.com/

https://www.mongodb.com/products/compass