Data Cleanup

REVIEW | HISTORY

Meets Specifications

Hey!

Great attempt made in the project! I appreciate your time and efforts on the same, especially the well-laid explanation and reasoning coupled with business logic and analytical skills. You are one step closer to graduating from the Nanodegree. 👏🏽

I have included certain remarks in the reviews in order to help you. I hope they'll be useful. If you need any help, feel free to post questions on https://knowledge.udacity.com/ Don't worry, we want the best from you; hence gave you all the possible areas of improvement.

If you are satisfied, please feel free to rate the review and provide your feedback. 😄

Wishing you all the best for future endeavours. Good luck 😄 👍🏽

Business and Data Understanding

✓ The section is written clearly and is concise. The section is written in less than 250 words.

1. What decisions need to be made?
   Management needs to decide on which city in Wyoming to open Pawdacity's newest store based on predicted yearly sales.

2. What data is needed to inform those decisions?
   The management requires the decision to be based on predicted yearly sales. Therefore to predict sales we shall require past sales data of all Pawdacity's stores, demographic data of cities in Wyoming, data on Wyoming population in various cities, and sales data of competitor stores.

✓ All the following questions have been accurately answered:

1. What decisions need to be made?
2. What data is needed to inform those decisions?

Q1 ✅
Awesome: You're right! The key business decision we need to make based on our analysis is which city should Pawdacity open its 14th store, based on yearly performance.

Q2 ✅
Awesome: You did a great job by listing down the variables useful for the analysis, since we're asked "What data is needed?"

Building the Training Set

✓ The averages for each column is correct in the training set

✅
Awesome: The averages obtained are accurate and precise for all columns.

Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*

Total Families	*62,653*	*5695.71*

Additional Resources

For more information on the importance of Data Cleaning, please refer *Data Wrangling- Lesson 2 Data Issues- Video Dirty Data*

✓ Outliers have been analyzed for each field in the training set.

The outliers are accurately identified.

The decision to keep, remove, or impute each outlier is well justified.

After pointing out all the possible outlier cities only one city should be decided upon to remove.

✅
Awesome: Nice work identifying the outliers- Gillette and Cheyenne. The reasoning to remove Cheyenne and keep others has been stated appropriately.

Because the city of **Cheyenne** has relatively massive outlier values across multiple fields it will lessen the model's ability to make predictions as keeping it in the model will skew all other predictions.
Therefore it would be best to filter out and remove the record of the city of **Cheyenne.**

The record of the city of **Rock springs** has an outlier value in the field of **land_area** with a value of 6620.20. Based on the fitted line on the scatter plot, the outlier is in line with the relationship, so I'd leave it in.

The record of the city of **Gillette** has an outlier value in field of **sales** with a value of 54,3132. Rather than removing the record of this city I will build a model with and without this record to compare the effect of the outlier.

Additional Information

Based on the IQR method results, you can either remove/keep Gillette or Cheyenne because-

1) Remove Gillette because it skews high in sales, yet does not skew relative to the other data fields in the training set and hence keep Cheyenne because it's inline with the linear relationship

2) Remove Cheyenne because it's unlike other cities for many fields (i.e. it's an outlier in many fields) and hence keep Gillette because the dataset is small and the values are close to Upper and Lower fence

DOWNLOAD PROJECT

RETURN TO PATH