Mane Street Horse and Pet:

Data Science and Business Analytics Internship

Jordyn Dolly

801273499

University of North Carolina - Charlotte

DSBA 6400

August 8th, 2023

Overview

Mane Street Horse and Pet is a family owned, local business in Waxhaw, North Carolina. Mane Street is one of the few local feed and farm supply stores in the area. Alison Bailey bought the store in 2010 with no formal background or training on how to run and operate a small retail business. Alison's background managing as a chief financial advisor for AAA of the Carolinas, being a CPA, VP of finance, CFO, and more helped prepare her to take own owning a small business. Since then, the business has grown to include a variety of pet feeds and supplies, but primarily focusing on horse and dog food. Mane Street Horse and Pet carries a variety of brands and products that are available for purchase in the store, in-store pickup, and delivery.

Since working at Mane Street Horse and Pet while working towards my master's in data science and business Analytics, the opportunity to show Alison how useful data science can be in helping to improve her business had arrived. While Alison has no formal data science or programming background, her experience in business analytics, human operations, and business operations have helped me become a more well-rounded future employee. Additionally, having the opportunity to delve deeper into the struggles and achievements that small businesses face has brought a greater appreciation for the constant work that is being done to remain successful. To not overwhelm Alison with data science jargon, I found that utilizing visualizations and graphs allowed her and I to have the most fruitful conversations on how sales may have been impacted in both a positive or negative way.

Data Collection

Data collection during the internship proved tedious and difficult. Mane Street Horse and Pet (MSHP) currently uses QuickBooks Desktop Point of Sale, an outdated and come October of 2023, unsupported POS system. The QuickBooks POS system has minimal reporting and data storage functions that are mainly being utilized to run reports of sales for customers in various rewards programs, and to run a report on inventory items to see what needs to be reordered. Additionally, Alison said she will run reports on the amount of a particular item that was sold within the past month and use that number to determine how much she will order for the next month.

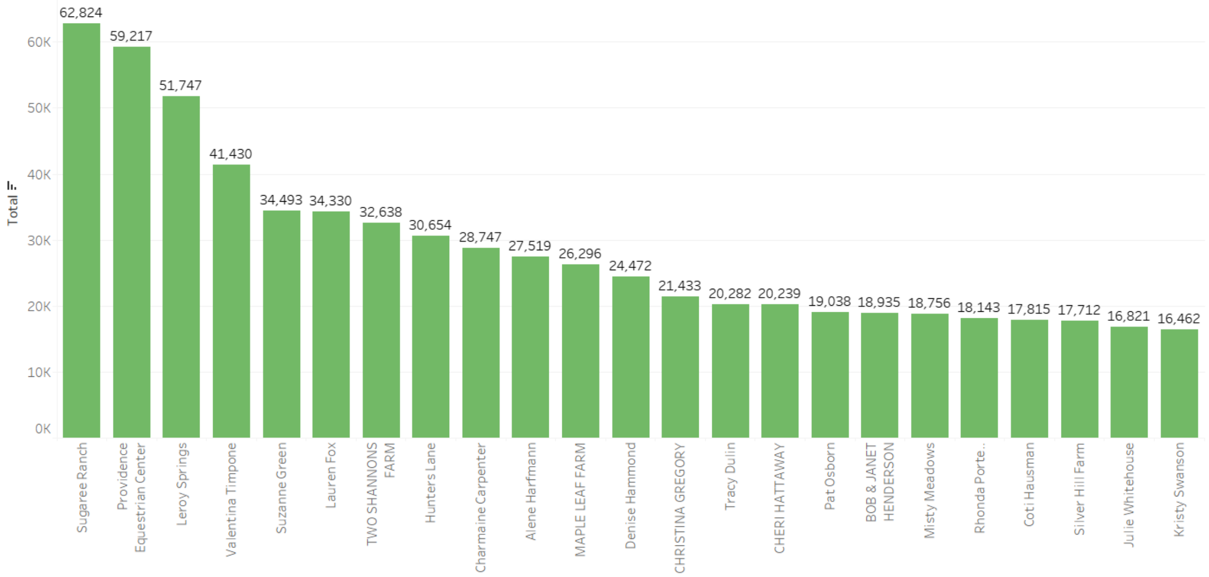
Most of the data was able to be pulled into excel files from QuickBooks POS. This was not a simple task, as many of the options were mislabeled and misleading. Due to the inability to have data storage in a cloud-based system, to gather the most up to date information I ended up with over four versions of the same file. Additionally, I had to manually correct each exported file for spacer columns and unnecessary headers. Gathering data that was not available in a "preset" excel spreadsheet then meant I had to copy items, transactions, customer information, dates, and more into an excel spreadsheet by hand. With over 6,000 card transactions being run each year, it was not an efficient use of my time to copy unnecessary data and instead for data like the "Highest Receipt Delivery Orders" file. To perform text analysis, I needed itemized receipts from customers and found there was no option to print multiple transactions with their itemized receipts. Therefore, I ran a report for the top sales within 2022 and 2023. From there I was able to select the top twenty valid delivery tickets, return to the individual customer history page and find their itemized receipt. Had the collection process for that data been more efficient, I would have been able to create more accurate text analysis models.

Utilizing Tableau

Tableau had proved to be extremely helpful in helping Alison visualize sales. The "Customer Sales Combined" file, was filtered to only include 2022 to find the customers who have purchased the most throughout the year. Alison and I agree that utilizing just 2022 would be the easiest to simplest way to look at sales after returning to normal operations after COVID-19. Each of these customers primarily focuses on horse boarding and training, resulting in anywhere from 10 to over 30 horses on each of these properties at any given time. Within Figure 1 Alison and I were most surprised to find that the top customer of the year was Sugaree Ranch. To provide some insight on this farm, they board and train horses year-round. This facility is not the most expensive in the area, nor do they travel to shows as frequently as other customers. Reflecting on this, many horse feeds can be found at various feed stores, so for customers who do frequently travel with their horses they are not always buying from MSHP. With Sugaree Ranch horses mainly stay on the property, and not traveling to shows this may be one of the largest contributing factors to Sugaree Ranch having the highest total sales in 2022.

Figure 1

Total Sales in 2022 by Customer

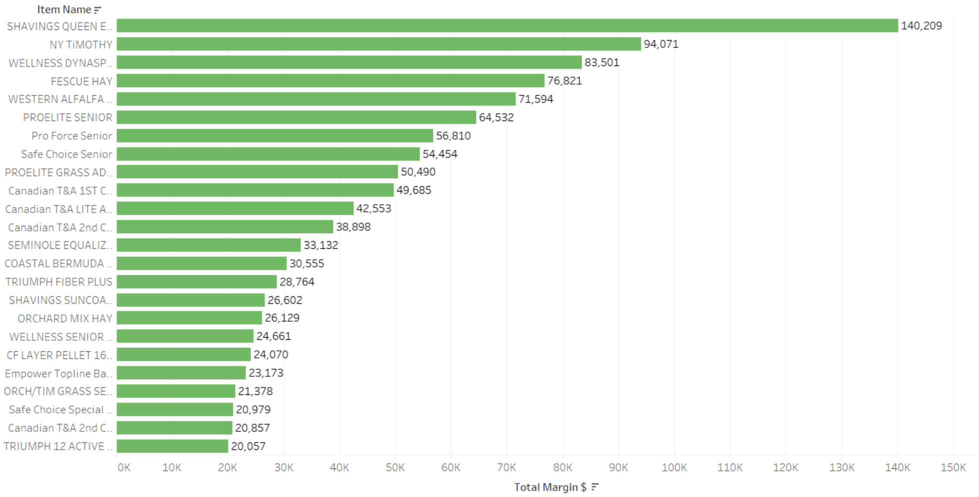


Continuing to look at Figure 1, these top customers have spent a small fortune on products at MSHP. While these customers are buying essentials for their animals like hay, bedding, and feed, customer feedback is rarely received. We believe that allowing these top customers to voice their opinions about products they would prefer to see in the store is extremely important. Not only does asking for their feedback show that MSHP values them as customers, but we may also find common requests among these customers that could improve sales.

Drilling down we wanted to look at the products that have sold the most over the past few years. Gaining further insight on the top selling products gives an insight into what customers are buying the most not only gives more insight on items that should have extra stocked, but also similar products and brands that customers may be interested in. Figure 2 illustrates the products who have brought the greatest amount of money, "Total Margin $". The top product, 'Shavings Queen Easy Sift', is a pine bedding that is most popular for horses but is commonly used for chickens and other animals. Additionally, various types of hay appear throughout the top products – NY timothy, fescue, western alfalfa, 1st cut timothy alfalfa, 2nd cut timothy alfalfa. While some of these products are only available seasonally, Figure 2 greatly illustrates where the greatest amount of money is being made.

Figure 2

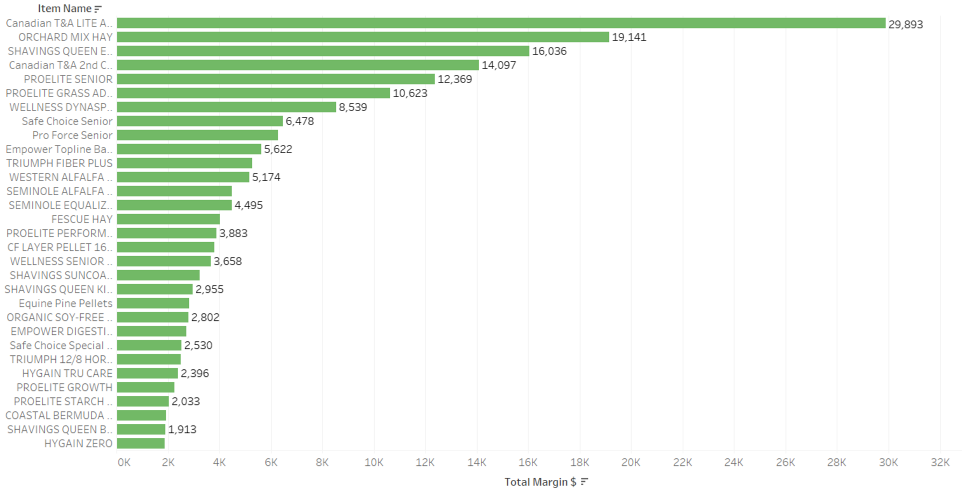
Top Items Sold from January 2018 to July 2023



While Figure 2 depicts the top selling items over multiple years, as product prices continue to rise, we can see a shift in the products currently trending to be the top selling products for 2023 in Figure 3. The relative positions of many of the products have changed, but hay and shavings remain as one of the top selling items. Additionally, we now see that the top selling horse feed is Nutrena's ProElite Senior, which is a direct competitor with Seminole's Wellness Dynasport. Both horse feeds are high quality products, but with a new Seminole regional sales representative not communicating efficiently and the Nutrena representative being very hands on and visiting various farms in the area, it is of little surprise that their products have grown in popularity.

Figure 3

Top Items Sold 2023 (January 1st to August 1st)



Another reason that we likely see hay and bedding as some of the top products sold amongst Figure 2 and Figure 3 is that hay and bedding are universal items. Quality hay and bedding are frequently bought by customers regardless of the feed being used. Many horse farms and owners are biased and have their preferred brand that they choose to feed their horses, however this is often not the case when it comes to hay or bedding. Instead hay and bedding are solely bought based on the quality of the items, price, and availability.

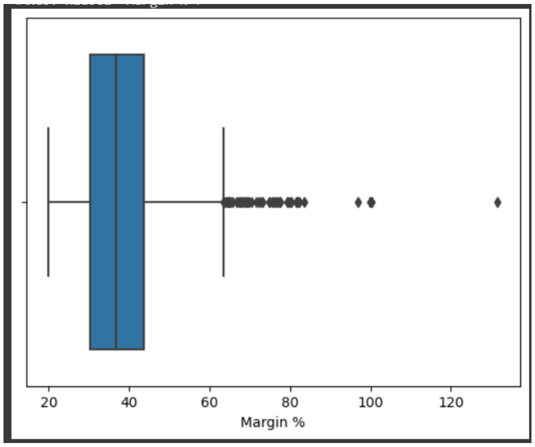
SVM for Minimum Margin Percentage

Introducing Alison to Support Vector Machines (SVM) was completely new to her. After explaining that utilizing SVM we can classify data, perform regression analysis, and detect outliers within the data she noted that detecting outliers in sales can be extremely useful. Frequently a customer will go out of town, try a new product, or a new customer will come into the store and purchase a product. While these are all situations where more product is being sold than on an average day, these outlier orders can result in extra product sitting and potentially going out of date.

Utilizing a box and whisker plot for the variable ‘Margin %’ in the ‘Qty Sold per Item 2018-2023’ data file, we can visualize the interquartile range and the outliers. In regard to margin percentage having an abnormally higher margin is not necessarily a bad thing, but if the margin is significantly higher than other products and they are not selling well, it may be worth reconsidering the margin percentage. Figure 4 shows that the median ‘Margin %’ falls slightly under 40% with the minimum at approximately 20% and the maximum around 63%. The figure does indicate that there are multiple outliers past the upper quartile, and it is likely that these are in store products like treats, fly sprays, buckets, and other barn and animal products.

Figure 4

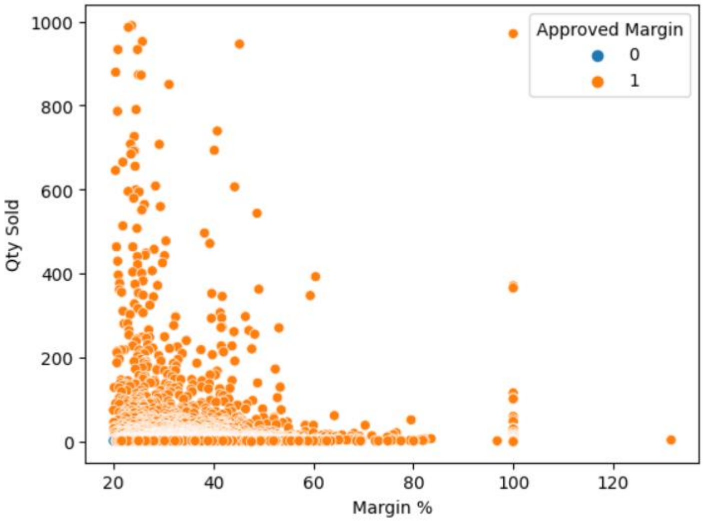
Box Plot of Margin % After Removing Quantity Sold Outliers



When I prompted Alison with the question “If you have to have one margin percentage for every product in the store, what would that margin percentage be?”, she responded that that margin percentage would be 20%. Figure 5 graphs the ‘Margin %’ and ‘Quantity Sold’ and is then coded to show a hue for the ‘Approved Margin’ which is set to 20% in this visualization. As we can see nearly all the data points are orange showing that they meet Alison’s approved margin value of 20%. However, it is possible to see there is at least one blue spot that indicates the margin value was less than 20%. With how low the quantity sold is for the one visible blue item, it is likely that there is a product that has not had its margin updated and has no sold between 2018 to 2023.

Figure 5

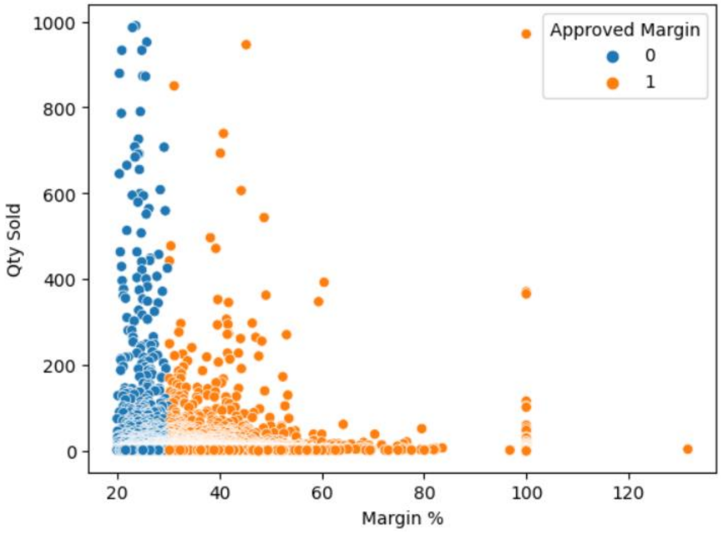
Margin % vs Quantity Sold: Colored by Margin Over 20.00%



To gain further insight on the range of items that fall within that 20% approved margin percentage, I increased the 'Approved Margin' value to 30%. In Figure 6 we can see that a large portion of the items above the 400-quantity sold level do not meet the approved margin of 30%. The reason we see a large drop off in the items that do not meet a minimum approved margin of thirty percent is because of horse feed. When MSHP must sell horse feed against big box stores likes Tractor Supply, makes it difficult for small businesses to make any money. These stores are buying significantly larger quantities of products at a time and therefore often get a discount given, frequently resulting in Tractor Supply selling their products at the cost that Mane Street Horse and Pet pays to bring it into the store. In finding the happy medium between making some money per bag of horse feed and keeping prices low enough to compete with big box stores, we can see that the margin percentage does have to be lower than other products.

Figure 6

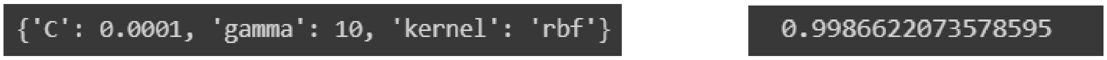
Margin % vs Quantity Sold: Colored by Margin Over 30.00%



Through the SVM analysis I was able to find the best fit parameters and grid score. The grid score reflects that there is near perfect accuracy at 0.99866. Utilizing SVM was more appropriate and was found to have better accuracy than the random forest model to classify whether the products met the approved margin percentage or not. In Figure 7 we can also see the optimal parameters for this data.

Figure 7

Best Parameters and Grid Score for SVM



```
{'C': 0.0001, 'gamma': 10, 'kernel': 'rbf'}
```

```
0.9986622073578595
```

Text Analysis of Top Delivery Orders

As a small business, MSHP does not have a huge following on social media. Alison has recently enlisted the help of an outside marketing team to help give the website a makeover and post more frequently on social media. While performing text and sentiment analysis on reviews left by customers, we found that text analysis on the products purchased may be insightful. One of the ways that MSHP can reach such a wide customer base and perform as well as they do is by running deliveries five days a week. With deliveries bringing in a bulk of the daily sales to the store, it was decided that using the top delivery sales tickets from 2022 and 2023 would provide enough information on the types of products being purchased. Utilizing the items on these delivery tickets as the text component, an intertopic distance map was created as shown in Figure 8.

From this figure we can see that topic 1 contains a combination of hay (‘alfalfa’, ‘hay’, ‘western’, ‘orchardgrass’, ‘coastal’), but mostly feed products (senior, proelite, empower, triumph, textured, select). These feed products mostly align with Nutrena branded items, with the ProElite Senior being the top selling feed product from Nutrena. Topic 1 most aligns with the most popular products sold. Topic 2 encompasses the next most popular items, like ration balancers. Ration balancers are products that are fed to horses in smaller quantities because they are denser in vitamins and minerals. Due to this these products often last longer and are purchased less frequently. Topic 3 shows some of the even lesser frequently bought items. We see ‘t/a’ or timothy-alfalfa hay occur in this topic and this is most likely due to the frequency at which hay is bought. Most farm managers buy enough hay to last them multiple weeks at a time, instead of buying a single bale every week. Many farms have their own connections and receive semi-truck loads full of hay at a time, so specialty hays are typically purchased in smaller quantities at MSHP. Looking at topic 4 we can see many non-feed or hay items. Things like salt blocks, Standlee brand hay cubes and pellets, and urine deodorizer pellets make up a large portion of this topic. Lastly, topic 5 contains the most niche products. Hygain being the largest part of this topic is a feed company based in Australia that is still working to gain popularity in the United States.

An additional way that we can take the ‘Highest Receipt Delivery Orders’ data and look at the similarity amongst deliveries is through a dendrogram. The dendrogram allows us to look at a hierarchical relationship amongst the delivery items and can then be put into separate clusters. Each of the lines on the dendrogram in Figure 9 illustrates the similarities between delivery orders. We can see the greatest number of similarities between order 0 and 8 and find a great number of similarities with orders 9, 10, 5, 17, and 12 as indicated by the blue line. From the dendrogram we see three different colors, representing three different topics as compared to the five topics selected in the distance map.

Figure 8

Intertopic Distance Map for Top Delivery Items

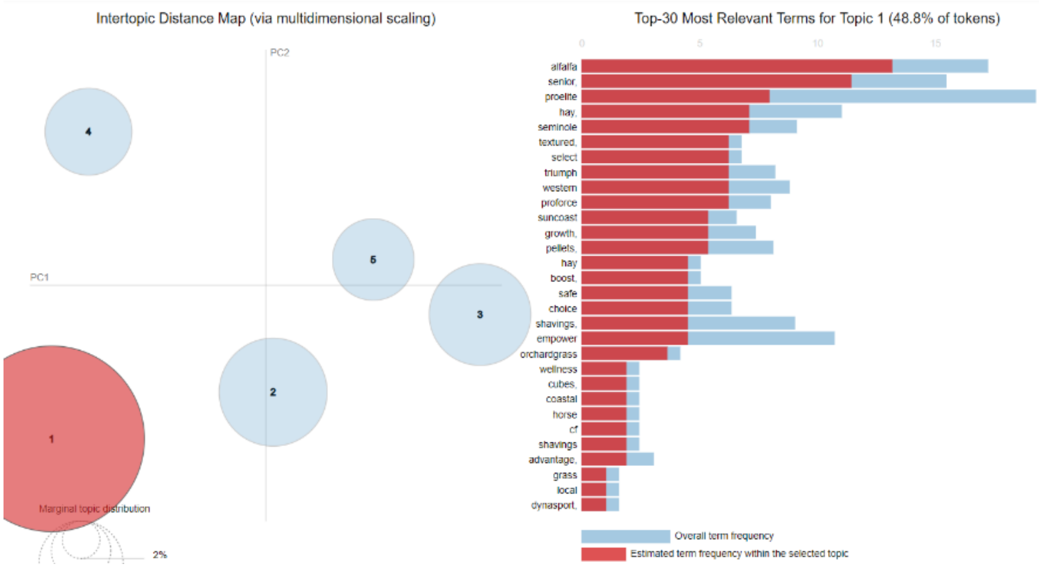
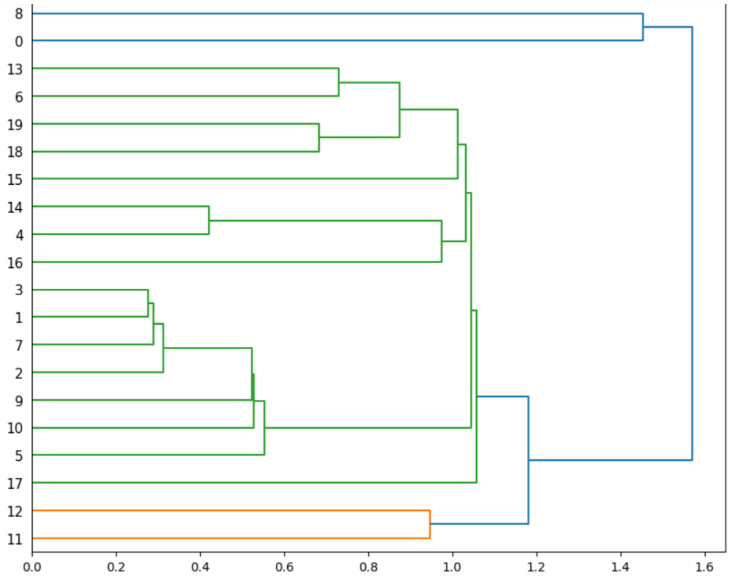


Figure 9

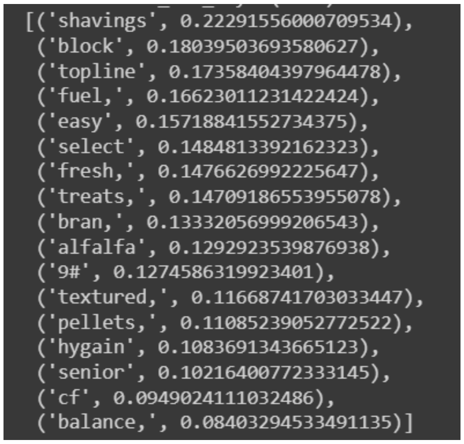
Dendrogram of Delivery Items from Top 20 Delivery Customers



Continuing to look at the usefulness of text analysis with the 'Highest Receipt Delivery Orders' data, we can calculate the percentage of likeliness that an item is purchased with another item. In Figure 10 we choose to use the word 'hay' amongst the text analysis and calculate the top relationships to that word. From the figure we can see that customers who purchase hay for a delivery order are also 22.3% likely to have shavings added to their delivery. When delivery orders are placed, whether it be over the phone, through email or over text, asking the question to these farms "Would you like to add any shavings to your delivery order?" could greatly increase sales. Without an order form for customers to look at when placing their order, many may not be aware of the types, sizes, brands, and prices of our shavings compared to where they may be currently purchasing from. Additionally, to not repeat the same question all the time, thinking about factors like the time of year and weather can help prompt other questions that will offer customers the chance to add more products to their deliveries. In small businesses, going the extra mile to ask if a customer needs additional items can greatly improve profit over the course of the year.

Figure 10

Relationship of the word "hay" throughout the Top Delivery Tickets through 2022 and 2023



```
[('shavings', 0.22291556000709534),
 ('block', 0.18039503693580627),
 ('topline', 0.17358404397964478),
 ('fuel,', 0.16623011231422424),
 ('easy', 0.15718841552734375),
 ('select', 0.1484813392162323),
 ('fresh,', 0.1476626992225647),
 ('treats,', 0.14709186553955078),
 ('bran,', 0.1333205699206543),
 ('alfalfa', 0.1292923539876938),
 ('9#', 0.1274586319923401),
 ('textured,', 0.11668741703033447),
 ('pellets,', 0.11085239052772522),
 ('hygain', 0.1083691343665123),
 ('senior', 0.10216400772333145),
 ('cf', 0.0949024111032486),
 ('balance,', 0.08403294533491135)]
```

Conclusion

Utilizing data science at Mane Street Horse and Pet has proved to be an educational experience for Alison but has also helped her gain greater insight on the relationships amongst products. Additionally, it has allowed me to use the tools and knowledge gained while pursuing my degree in data science and business analytics. Despite facing issues with exporting data and the files not including all information, the use of Tableau, SVM, and text analysis have proved to be the most insightful tools. I was able to gain experience with the behind the scenes work that goes on to keep a small business operating.

Appendix A

Text Analysis: Responsible for figures 8,9,10

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_excel('Highest Receipt Delivery Orders.xlsx')
df

import gensim
from gensim import corpora, models

docs = df['Items Purchased']
docs

corpus = [doc.split() for doc in docs]
corpus

dictionary = corpora.Dictionary(corpus)

DFM = [dictionary.doc2bow(doc) for doc in corpus]

term_maps = dictionary.token2id

term_maps = {v: k for k, v in term_maps.items()}
term_maps

myDFM = pd.DataFrame(
    gensim.matutils.corpus2csc(DFM).T.toarray()).rename(columns=term_maps)

myDFM

myDFM.sum().sort_values(ascending=False)

df['Cleaned Text'] = df['Items Purchased'].str.replace(r'[^\w\s]+','')

df['Cleaned Text']

import nltk
nltk.download('stopwords')

from nltk.corpus import stopwords
stop = stopwords.words('english')
df['Cleaned Text']=df['Cleaned Text'].apply(
    lambda x: " " .join(x for x in x.split() if x not in stop))

df['Cleaned Text']

df['Cleaned Text']=df['Cleaned Text'].apply(
    lambda x: " ".join(x.lower() for x in x.split()))

df['Cleaned Text']

docs_clean = df['Cleaned Text']
corpus_clean = [doc.split() for doc in docs_clean]
dictionary_clean = corpora.Dictionary(corpus_clean)
DFM_clean=[dictionary_clean.doc2bow(doc) for doc in corpus_clean]

term_maps = dictionary_clean.token2id
term_maps = {v:k for k, v in term_maps.items()}
term_maps

myDFM_clean = pd.DataFrame(
    gensim.matutils.corpus2csc(DFM_clean).T.toarray()).rename(columns=term_maps)

myDFM_clean

tfidf = gensim.models.TfidfModel(DFM_clean)
DFM_tfidf=tfidf[DFM_clean]

myDFM_clean = pd.DataFrame(
gensim.matutils.corpus2csc(DFM_clean).T.toarray()).rename(columns
= term_maps)
myDFM_clean

SVD_model = gensim.models.LsiModel(DFM_tfidf,
                                   id2word = dictionary_clean,
                                   num_topics = 10)
SVD=SVD_model[DFM_tfidf]
SVD_result = pd.DataFrame(gensim.matutils.corpus2csc(SVD).T.toarray())
SVD_result

"""Topic Modeling"""

n_topics = 4
ldamodel = gensim.models.LdaModel(DFM_clean,
                                  num_topics=n_topics,
                                  id2word = dictionary_clean,
                                  passes=20)
```

```
!pip install pyLDAvis
import pyLDAvis
pyLDAvis.enable_notebook()

import pyLDAvis.gensim_models
vis = pyLDAvis.gensim_models.prepare(ldamodel,DFM_clean,dictionary_clean)
vis

"""Similarity and Clustering"""

from gensim.similarities import MatrixSimilarity
from scipy.cluster import hierarchy
import matplotlib.pyplot as plt

index = MatrixSimilarity(DFM_clean,
                         num_features=len(dictionary_clean))
distance = 1-index[DFM_clean]

Z = hierarchy.linkage(distance, 'single')
plt.figure(figsize=(15,10))
dn = hierarchy.dendrogram(Z, orientation='right',leaf_font_size=11)

text_sim = pd.DataFrame(index[DFM_clean])
text_sim[0].sort_values(ascending=False)

from gensim.models import Word2Vec
model = Word2Vec(corpus_clean,min_count=1)

sim=model.wv.most_similar('hay',topn=17)
sim
```

Appendix B

SVM Analysis: Responsible for figures 4,5,6,7

```
import pandas as pd
df = pd.read_excel('Qty Sold per Item 2018 - 2023.xlsx')

df.describe()

df.isnull().sum()/len(df)

df_cleaner = df.query('`Qty Sold` <= 1000')
```

```
df_cleaned = df_cleaner.query('`Margin %` >= 20.0')
df_cleaned

df_cleaned = df_cleaned.drop(columns=['Vendor', 'Item Name'])

def Approved_Margin(x):
  if x>20:
    return 1
  else:
    return 0

df_cleaned['Approved Margin']=df_cleaned['Margin %'].apply(Approved_Margin)

import seaborn as sns
sns.boxplot(x=df_cleaned["Qty Sold"])

from matplotlib import pyplot as plt
sns.scatterplot(data=df_cleaned,
                x='Margin %',
                y='Qty Sold',
                hue='Approved Margin')
plt.show()

y = df_cleaned['Approved Margin']
x = df_cleaned.drop('Item #', axis=1)

from sklearn.preprocessing import StandardScaler
x = StandardScaler().fit_transform(x)

import numpy as np
print(np.mean(x,axis=0))
print(np.std(x,axis=0))

from sklearn.model_selection import train_test_split

x_train,x_val,y_train,y_val = train_test_split(x,y,
                                               test_size=0.3,
                                               random_state=0)

from sklearn import svm

linearSVM = svm.SVC(kernel='linear')
linearSVM.fit(x_train,y_train)

linearSVM.C
linearSVM.score(x_val,y_val)

radialSVM = svm.SVC(kernel='rbf')
radialSVM.fit(x_train,y_train)

radialSVM.score(x_val,y_val)

from sklearn.model_selection import GridSearchCV

param = {'C':[0.1,0.5,1,5,10],
         'gamma':[1,0.1,0.01,0.001],
         'kernel':['rbf','linear']}

SVM = svm.SVC()

grid=GridSearchCV(estimator=SVM,
                  param_grid=param,
                  verbose=3,
                  cv=10)

grid.fit(x_train, y_train)

grid.best_params_

grid.score(x_val,y_val)
```

Appendix C

Regression Analysis

```
import pandas as pd
import numpy as np

df = pd.read_excel('Qty Sold per Item 2018 - 2023.xlsx')
df

df.info()

df.isnull().sum()/len(df)

df['Margin %']

df['Margin %'].value_counts()
```

```
df = df.dropna()
df.isnull().sum()/len(df)

margin_over20 = df['Margin %'] >= 20.00
margin_over20.value_counts()

NA_names=['Qty Sold','Ext Price','Ext Cost','Margin %','Total Margin $',
'Vendor', 'Item #']
df[NA_names]= df[NA_names].fillna(df[NA_names].mean())
df.describe(include='all')

dummy_data = pd.get_dummies(df[['Vendor']], drop_first=True)
dummy_data = dummy_data[['Vendor_CPC Carolina Distributing','Vendor_FLORIDA
HARDWARE','Vendor_Nutrena WXW','Vendor_QUEEN WOOD PRODUCTS','Vendor_RJ Matthews
Company','Vendor_Seminole Feed','Vendor_Southeast Pet','Vendor_Long Meadows
Farm','Vendor_M A Nance Farms','Vendor_Langridge Supply', 'Vendor_Peter
Destefano']]
dummy_data = dummy_data.dropna()
dummy_data.describe()

num_names = ['Qty Sold','Ext Price','Ext Cost','Margin %','Total Margin $']
X = pd.concat([df[num_names],dummy_data], axis=1)
X.describe(include='all')

X.skew(axis=0, skipna=True)
#everything is highly skewed so need a log transformation

X.hist(bins=50, figsize=(15,15))

def combine_Margin(x):
  if x>0:
    return 1
  else:
    return 0

X['Margin %']=X['Margin %'].apply(combine_Margin)

X.isnull().sum()/len(df)

X['Ext Price']=np.log10(X['Ext Price'])

X.corr()

X=X.drop(columns=['Ext Cost'])
```

```
y=df['Qty Sold']

from sklearn.model_selection import train_test_split

X_train,X_val,y_train,y_val=train_test_split(X,y,
                                             test_size=0.3,
                                             random_state=0)

import statsmodels.api as sm

log_reg = sm.Logit(y_train,X_train).fit()
print(log_reg.summary())

from sklearn.metrics import confusion_matrix

prediction_prob = log_reg.predict(X_val)

prediction = list(map(round,prediction_prob))
confusion_matrix(y_val,prediction)

from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from matplotlib import pyplot as plt

lr_auc = roc_auc_score(y_val, prediction_prob)
print('Logistic: ROC AUC =%.3f' % (lr_auc))

lr_fpr, lr_tpr, _ = roc_curve(y_val,prediction_prob)

plt.plot(lr_fpr, lr_tpr, marker='.')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.show()
```