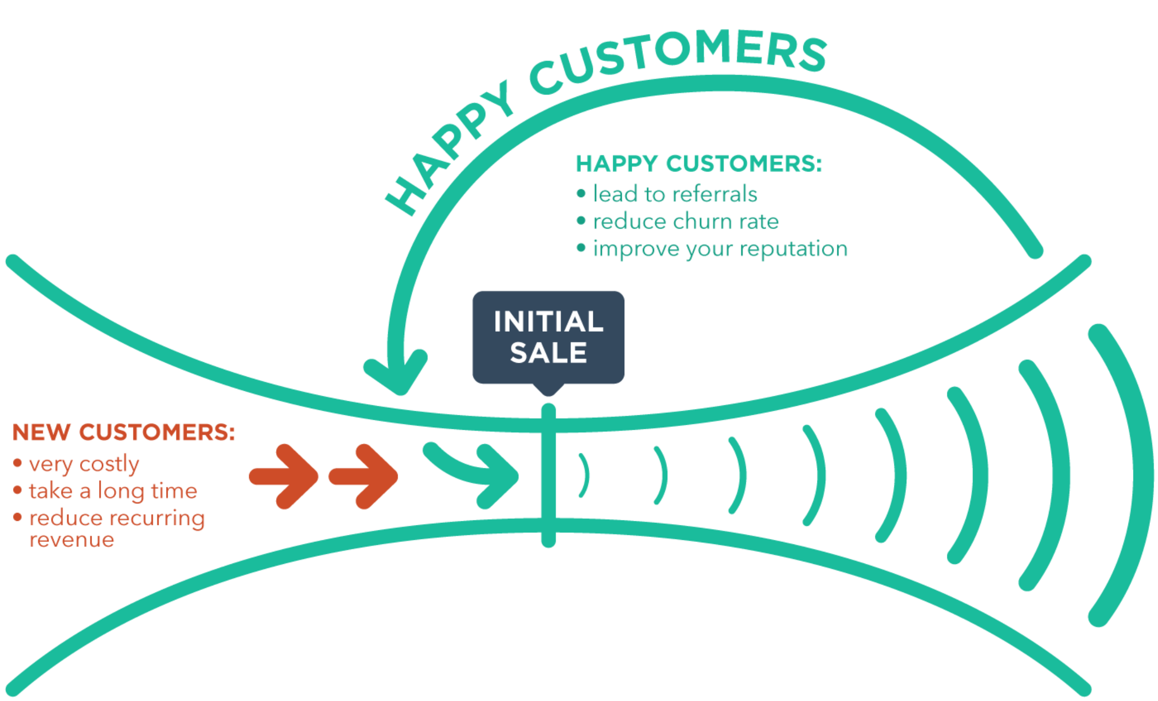
Data Capstone 1

Eric Huynh

April 2 Cohort

The Problem & The Goal

- Poorly stocked inventory can cause poor customer experience.
  - Causes increased churn rates and loss in profits.
- Production efficiency is limited by the lack of knowledge on product performance.
- **Goal:** Provide a visual analysis of the data and generate a predictive model that can forecast sales

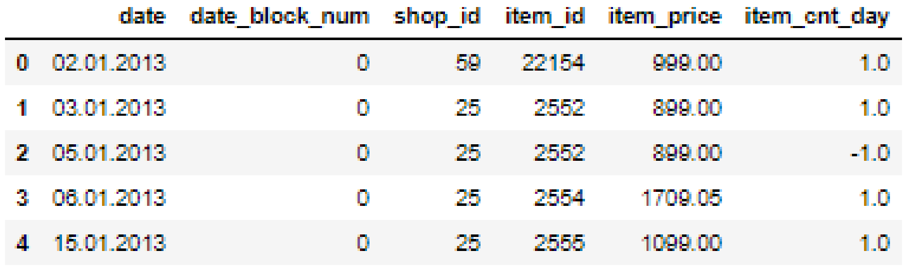


The Clients

- The model will provide a reliable method for the client to forecast the performance of their products for up to a month
- Individual Retail Stores:
    - Knowledge of popular items allows for retail stores to constantly update their shelves with the best products to potentially bring in larger profit margins
    - Managers will also be able to reliably predict efflux of each product which will help maintain properly stocked inventory
- Corporate Retail Stores:
    - Will utilize the model for the same purposes as individual retail stores
    - However, poor customer experience here will have a larger effect as it affects customer loyalty and reception of corporate reliability
- Production Companies:
    - Can help prioritize the manufacturing process to focus efforts on items of higher sale value.
    - If expanded with geographical information, production companies can extend their sales to better performing areas

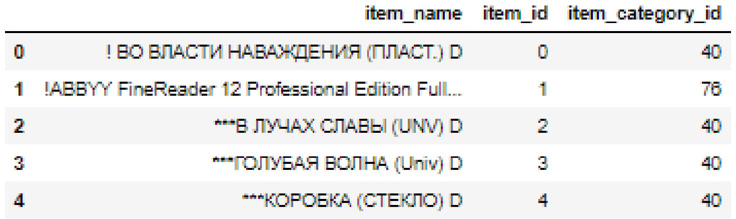
The Data

- The training data consisted of 6 variables and nearly 3 million data entries ranging from Jan. 2013 to Oct. 2015.
  - The original date format was DD.MM.YYYY.
- The items data consisted of information on 22,170 different items along with their appropriate item categories.
- The item categories data and shops data each had 84 and 60 entries respectively.



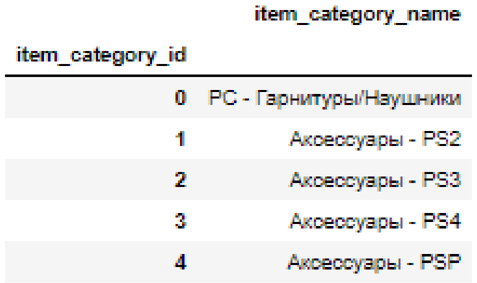
	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
0	02.01.2013	0	59	22154	999.00	1.0
1	03.01.2013	0	25	2552	899.00	1.0
2	05.01.2013	0	25	2552	899.00	-1.0
3	06.01.2013	0	25	2554	1709.05	1.0
4	15.01.2013	0	25	2555	1099.00	1.0

Figure 1. Quick look at provided training dataset



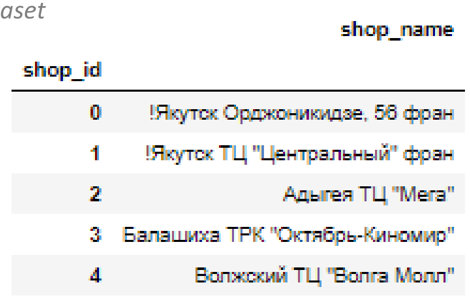
	item_name	item_id	item_category_id
0	! ВО ВЛАСТИ НАВАЖДЕНИЯ (ПЛАСТ.) D	0	40
1	!ABBYY FineReader 12 Professional Edition Full...	1	76
2	***В ЛУЧАХ СЛАВЫ (UNV) D	2	40
3	***ГОЛУБАЯ ВОЛНА (Univ) D	3	40
4	***КОРОБКА (СТЕКЛО) D	4	40

Figure 2. Quick look at provided items dataset



	item_category_name
item_category_id	
0	PC - Гарнитуры/Наушники
1	Аксессуары - PS2
2	Аксессуары - PS3
3	Аксессуары - PS4
4	Аксессуары - PSP

Figure 3. Quick look at provided item categories dataset

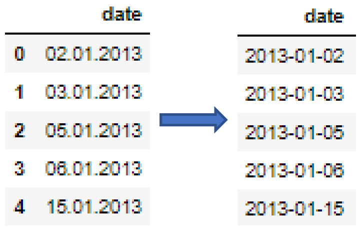


	shop_name
shop_id	
0	!Якутск Орджоникидзе, 56 фран
1	!Якутск ТЦ "Центральный" фран
2	Адыгея ТЦ "Мега"
3	Балашиха ТРК "Октябрь-Киномир"
4	Волжский ТЦ "Волга Молл"

Figure 4. Quick look at provided shops dataset

The Cleanup

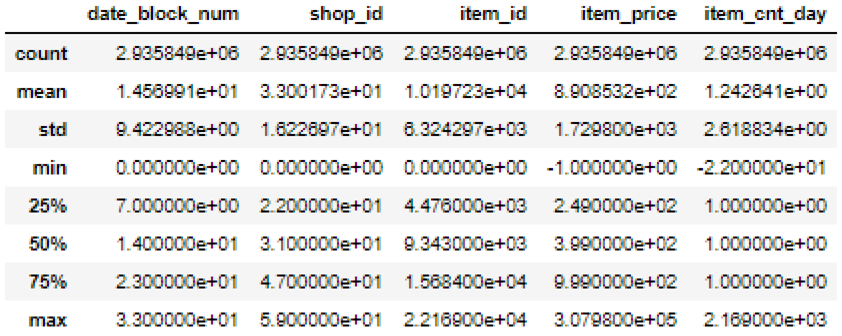
- The string format of the date column was changed to a function friendly format (DD.MM.YYYY -> YYYY-MM-DD) for conversion to datetime objects.
- There were two extreme values in the item_price variable that do not make practical sense and were removed from the dataset.
  - There was an outlier on the higher price spectrum at 307,980 which is magnitudes than the second greatest at 59,200.
  - There was also single negative outlier at -1.
- There were also 7300 entries that have negative values in the item_cnt_day variable which could represent returns to a store. However, because no additional information regarding the data could be gathered, I chose not to remove them.



	date
0	02.01.2013
1	03.01.2013
2	05.01.2013
3	06.01.2013
4	15.01.2013

date
2013-01-02
2013-01-03
2013-01-05
2013-01-06
2013-01-15

Figure 5. String format conversion to date parseable format

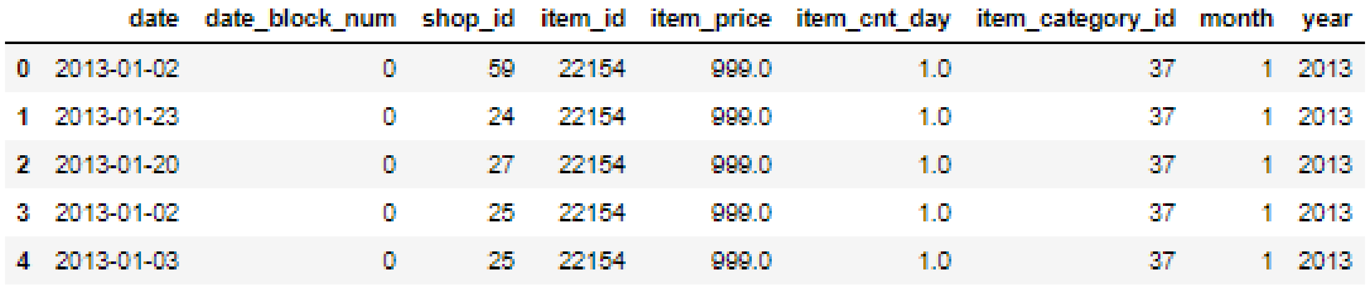


	date_block_num	shop_id	item_id	item_price	item_cnt_day
count	2.935849e+06	2.935849e+06	2.935849e+06	2.935849e+06	2.935849e+06
mean	1.456991e+01	3.300173e+01	1.019723e+04	8.908532e+02	1.242641e+00
std	9.422988e+00	1.622697e+01	6.324297e+03	1.729800e+03	2.618834e+00
min	0.000000e+00	0.000000e+00	0.000000e+00	-1.000000e+00	-2.200000e+01
25%	7.000000e+00	2.200000e+01	4.476000e+03	2.490000e+02	1.000000e+00
50%	1.400000e+01	3.100000e+01	9.343000e+03	3.990000e+02	1.000000e+00
75%	2.300000e+01	4.700000e+01	1.568400e+04	9.990000e+02	1.000000e+00
max	3.300000e+01	5.900000e+01	2.216900e+04	3.079800e+05	2.169000e+03

Figure 6. Note the positive and negative outliers in the item_price column

The Transformation

- The name columns in the three datasets (items, shops, and item categories) had a mix of Russian and English.
  - Not time efficient to extract only the English characters.
  - Costly to translate the Russian characters.
  - Decided to remove the names from my analysis
    - Valid choice as each item, shop, and item category had a unique ID.
- As a result, only the item dataset needed to be merged with the training dataset (for the unique item category IDs).
- Additional month and year columns were constructed from the date column for use as categorical variables.



	date	date_block_num	shop_id	item_id	item_price	item_cnt_day	item_category_id	month	year
0	2013-01-02	0	59	22154	999.0	1.0	37	1	2013
1	2013-01-23	0	24	22154	999.0	1.0	37	1	2013
2	2013-01-20	0	27	22154	999.0	1.0	37	1	2013
3	2013-01-02	0	25	22154	999.0	1.0	37	1	2013
4	2013-01-03	0	25	22154	999.0	1.0	37	1	2013

Figure 7. Transformed training data with all the necessary variables for exploratory data analysis.

The Trends

- Time-series line plot of 3 randomly selected shops from January 2013 to October 2015 show a general decline in sales over the years (2013 > 2014 > 2015).
    - Additionally, there is a visible peak on December of 2013 and 2014 which may due to some sort of cyclical trend.
- Applied a combination of a histogram and a barplot to emphasize the cyclical trends
    - There are increased number of entries and total sales in December for years 2013 and 2014.

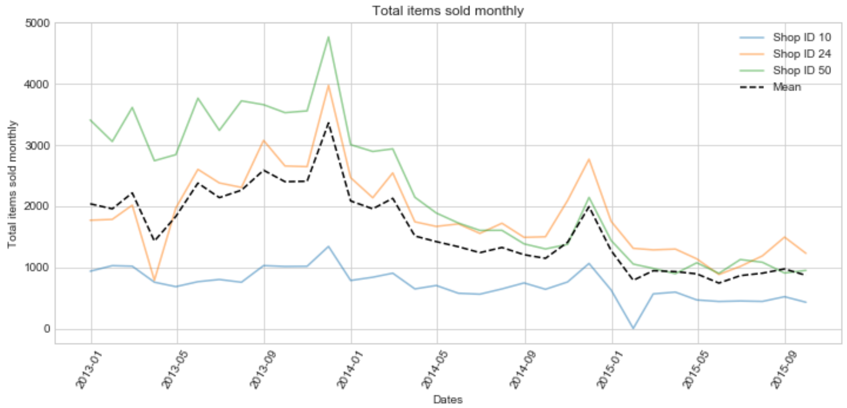


Figure 8. The time-series plot shows an apparent decline in overall sales over time.

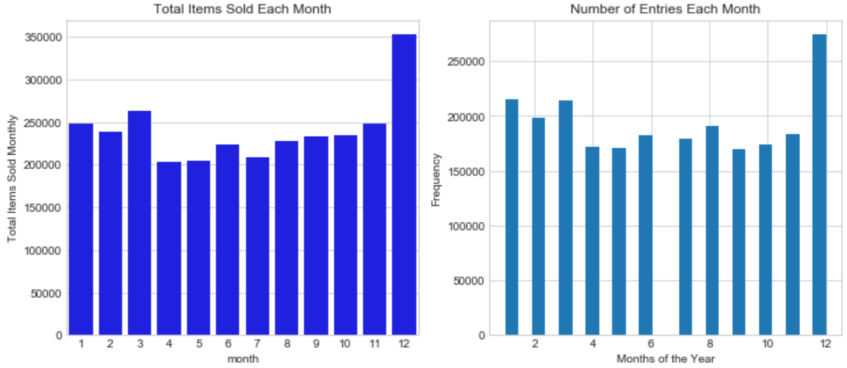


Figure 9. The histogram and bar plot of data from 2013 and 2014 show increased total entries and total sales during the month of December.

The Statistics

- In order to make the development and deployment of the model more computationally feasible, groups were made for originally just **item category IDs**, **shop IDs**, and later **item IDs**.
    - 17 Groups for Item Categories IDs
    - 13 Groups for Shop IDs
    - 24 Groups for Item IDs
- Groups were statistically validated to be different from each other via two-sample t-test with p-value 0.05.

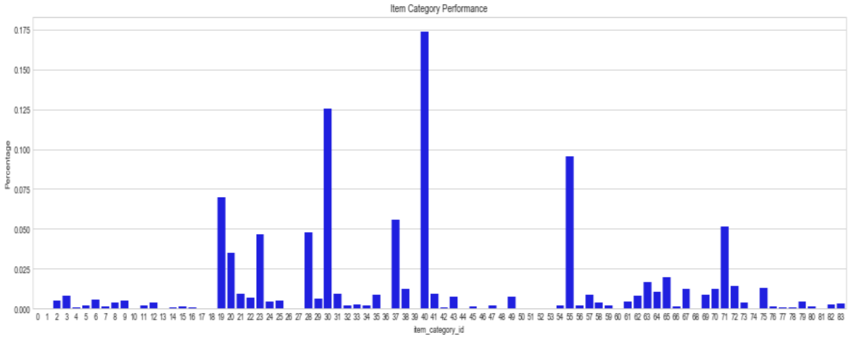


Figure 10. The normalized bar plot shows performance of shops relative to each other.

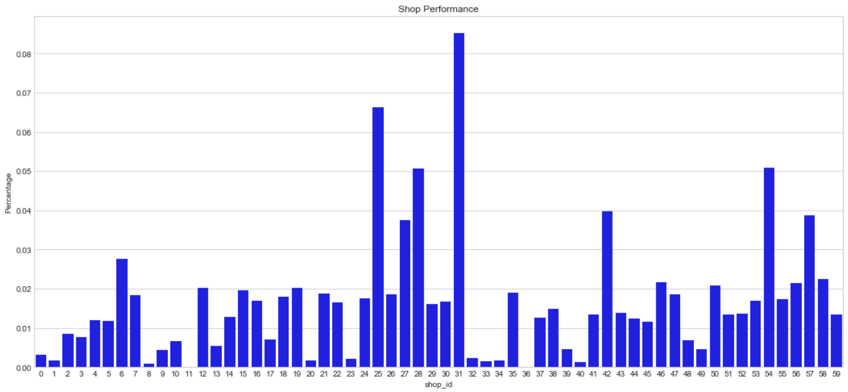


Figure 11. The normalized bar plot of item categories shows relative performance.

The Model

- Simple Linear Regression model Using 4 Variables:
  - Item Category IDs: Both the grouped and independent versions of the Item Category IDs were used in the development of the model
  - Shop IDs: Both the grouped and independent versions of the shop IDs were used in the development of the model
  - Month & Year: Categorical representations of the previous date variable used in the time-series plot
- Initial deployment of the model performed poorly. As a result, different combinations of the variables (besides Month & Year) were attempted.
  - In total, there were 5 tested and deployed models.

The Results

- Model 1, 3, 4, 5:
  - Predictions consistently underestimate the actual values.
  - Irregular max predicted value cutoff not reflected in actual values
  - Contains negative predicted values along with a right skew in predictions

- Model 2:
  - Similar trends to all other models
  - Model 2 does not have the irregular max predicted value cutoff in the predictions for the split test data

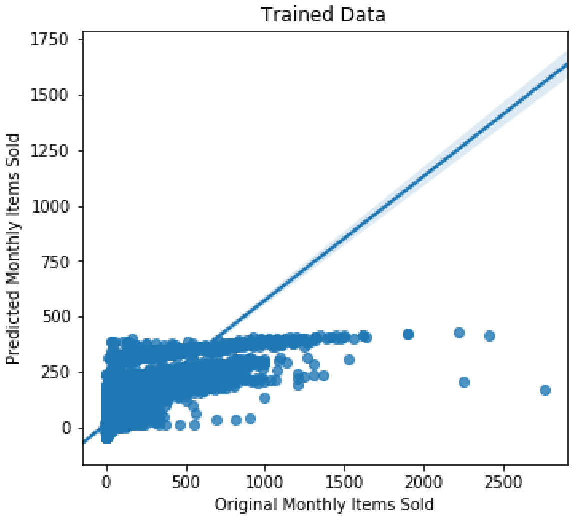


Figure 12. Regression plot of model 1

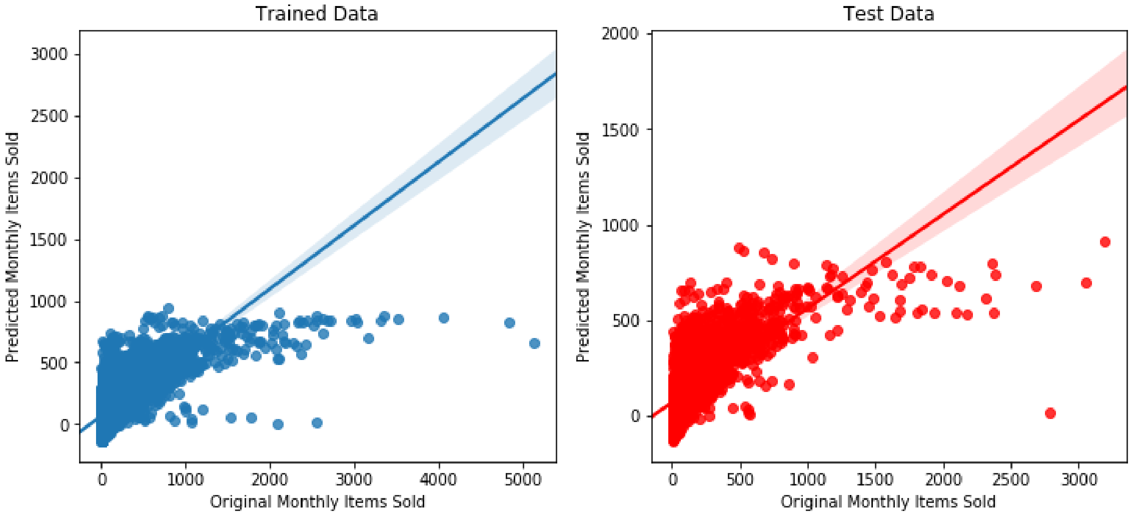
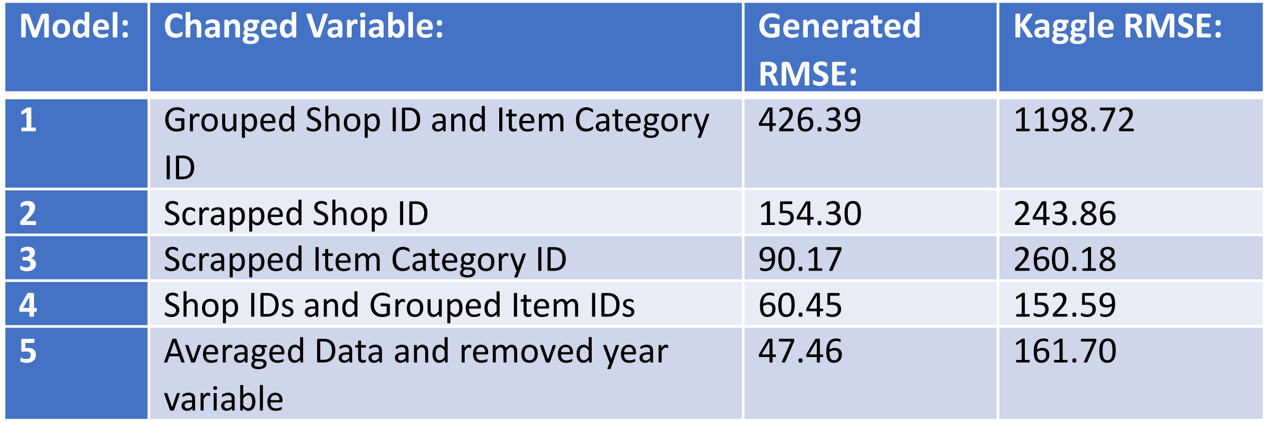


Figure 13. Regression plot of model 2

The Scores

- Variables used in each model. ***Note:** All models used manually inputted values for month and year to reflect Nov. 2015
- **Model 1:** Grouped Shop ID and Grouped Item Category ID
- **Model 2:** Shop IDs and Grouped Item Category ID
- **Model 3:** Shop IDs and Item Category IDs
- **Model 4:** Shop IDs and Grouped Item IDs
- **Model 5:** Shop IDs and Group Item IDs. Averaged the data across the months and removed the year



Model:	Changed Variable:	Generated RMSE:	Kaggle RMSE:
1	Grouped Shop ID and Item Category ID	426.39	1198.72
2	Scrapped Shop ID	154.30	243.86
3	Scrapped Item Category ID	90.17	260.18
4	Shop IDs and Grouped Item IDs	60.45	152.59
5	Averaged Data and removed year variable	47.46	161.70

Figure 14. All models trained and deployed along with their respective scores. Also notes the variable changes from the immediately previous model.