Grocery Store Analysis

Siddharth Kushawaha

Aishwarya Shastry

Harsh Shah

James Spaniak

Stevens Institute of Technology

"We pledge our honor to abide by the Stevens Honor System."

Abstract

The purpose of this project is to build an efficient sales modeling system for a retail organization using prior sales data and additional datasets provided. We intend to utilize statistical techniques to understand the relation between oil prices, unit sales, transactions, Store Type, and Store Locale.

List of Variables

- Id - unique identifier for dataset, [each product id] does not affect analysis.
- Type (Qualitative) – Types of Store (Classified by the Store itself) A, B,C,D,E
- Locale (Qualitative) – National, Regional, Local
- Dcoilwtico (Quantitative) – Daily oil prices
- Unit Sales (Quantitative) – Total count of product sold.
- Transactions (Quantitative) – Daily Transactions (Sum of Products sold everyday) [Target Variable]

Table of Contents

Introduction....4

Preliminary Investigation of Statistical Measures of Dataset Features 4

Initial Visualizations and Assumptions Regarding Variable Behavior 5

Determining the Distribution of Each Random Variable 9

Central Limit Theorem Exploration: Investigating Transactions 10

Constructing a Confidence Interval for Transactions from Random Samples 12

Hypothesis Testing Using the Same Random Transactions Sample 13

Comparing Different Datasets: Type and Locale with Transactions....14

Model Implementation and Results 15

Conclusion 19

Introduction

In order to observe how the theoretical basis of important statistical concepts influences the dynamics of various real-world applications, an analysis was conducted of a given Grocery Stores dataset encompassing n=100000 random daily sampled values for six variables of interest related to the Transactions and key potential influencers thereof: Date, Id, Type, Locale, Unit Sales and Dcoilwtico.

In this work, histograms and Q-Q plots were used to study the distribution of each variable, with Shapiro-Wilk and K-S testing as supplementary tools to reach a determination. These visualization methods were also used to confirm the principles of the Central Limit Theorem (CLT) and explore its consequences in this context. Confidence intervals were constructed to determine whether they were viable captures of the population mean of the ETF variable. The analysis tested our knowledge of hypothesis testing to reach correct conclusions about the mean and standard deviation of Transactions, as well as when comparing the Dcoilwtico and Unit Sales variables. Regression analysis and model assessment were also performed by fitting a line to the Transactions and interpreting the results of the scatterplot, histogram, and Q-Q plot output of the residuals.

In this work, we observe that the low correlation among the variables poses challenges to the predictive capabilities of linear regression models, and evidence of such is presented throughout. Also, our qualitative variables have multiple unique attributes leading to increasing the number of dimensions by using One Hot Encoding. There is, however, not enough evidence to refute normality assumptions, or that the data contradicts the CLT in any way.

Preliminary Investigation of Statistical Measures of Dataset Features

The investigation of the data started with a calculation of the total sample count, sample mean, sample standard deviation, minimum, 25^{th} percentile, median, 75^{th} percentile, and maximum sample value of each random variable. Notably, the standard deviation for the Transactions is very high suggesting that there is huge variance in the dataset. Even the mean values are very less in comparison to the maximum value suggesting huge right tail-skewness in the dataset. Similarly, unit prices also seem to right skewed data.

	Id	Unit Sales	Dcoilwtico	Transactions
count	1.000000e+05	100000.000000	100000.000000	100000.000000
mean	3.117355e+07	4.710928	85.789144	1958.454940
std	5.367097e+06	3.659894	19.744334	1143.012388
min	2.165766e+07	0.069000	53.450000	422.000000
25%	2.624134e+07	2.000000	65.940000	1181.000000
50%	3.255332e+07	4.000000	88.890000	1625.000000
75%	3.640353e+07	7.000000	104.760000	2519.000000
max	3.859423e+07	15.000000	107.950000	8120.000000

Figure 1: Data Description

Let us deep dive and check correlations between the available features. The sample correlations among each pair of the random variables were obtained. All pairs except for those that examined the variable against itself produced a correlation of close to zero, indicating that a strong linear relationship in either the positive or negative direction is not present. Transactions and Type is showing negative but not significant correlation.

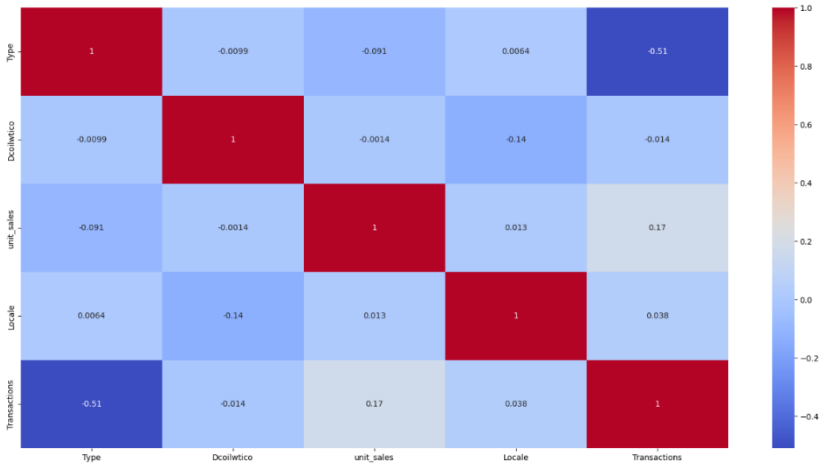


Figure 2: Feature Correlation Heatmap among random variable pairs.

Initial Visualizations and Assumptions Regarding Variable Behavior

Various visualization tools including histograms, time series graphs, and scatterplots were formed for the purpose of extracting general information about individual feature

distribution and behavior over time, as well as demonstrating graphically the relationships between the ETF column and the oil, gold, and JPM columns.

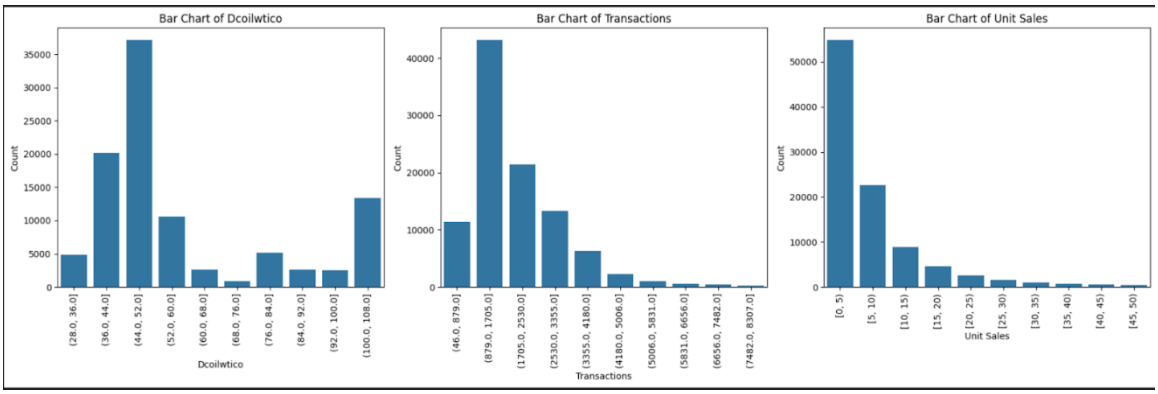


Fig 3: Distributions of each quantitative variable

- Dcoilwtico: The graph shows the distribution of oil prices in $10 increments and each bar represents the counts of the oil price with specific price range. This is a right skewed meaning lower oil prices are more common.

- Transaction: The graph shows distribution in 10 transactions each. Highlights the frequency of transaction volumes. This is also rightly skewed, meaning that the data with fewer transactions are more common than the days with a very high number of transactions.

- Unit Sales: The data is binned into intervals of 5 units. The adjustment is made to give a clear view of the sales volume ranges, for better clarity. This is also right skewed, most of the sales counts are low with infrequent occurrences of very high sales counts.

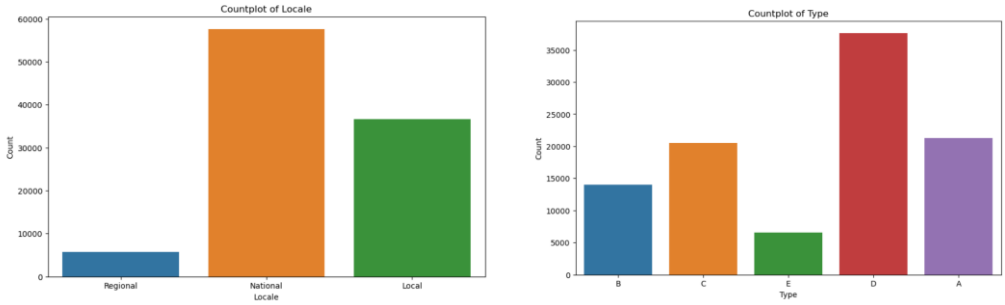


Fig 4: Count plot of Qualitative Random Variables in the dataset

From Figure 4, we can conclude that major Transactions were related to National Locale than other attributes. Similarly, D type of stores is contributing more towards the total revenue of the Grocery Chain.

Fig 5: Pairplot of the dataset

Scatter plots comparing variables. Along the distribution of value counts and off the diagonal are the variables scattered plotted against each other. The locale vs type scatter plots are a 5x3 set of points. Where the continuous variables are a standard scatter plot where a regression line could be drawn. Finally, the categorical variables vs the continuous variables can show the variety of data points within each category. With the large number of data points and high number of variances within the continuous variables it is hard to determine relationships visually. Although we can see cutoffs such as 0 for unit sales, where there is only a single return observation. Additionally, we can see where the main grouping of values lies.

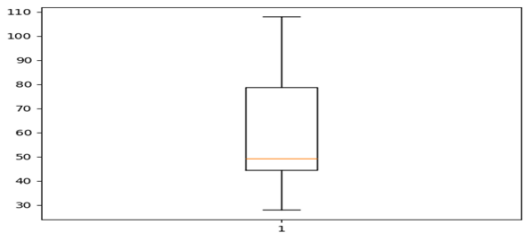


Fig 6: Oil Price Boxplot

Oil does not have any outliers outside of IQR*1.5. The median is clearly shifted downwards.

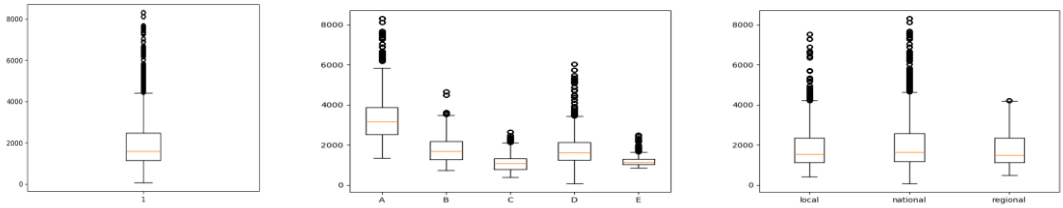


Fig 7: Transactions Boxplot - Overall and Categorical Breakdown (Type & Locale)

Transactions contain a large number of outliers past 1.5*IQR range and should be looked at to see how it affects our model. Large variance in transactions as the 35-75 IQR is over 1000 wide. The median is skewed down with outliers in the large positive value range. The IQR size seems to be the same with number outliers varying by type. For type, A and D have some larger number of outliers and larger IQR. The median differs per type.

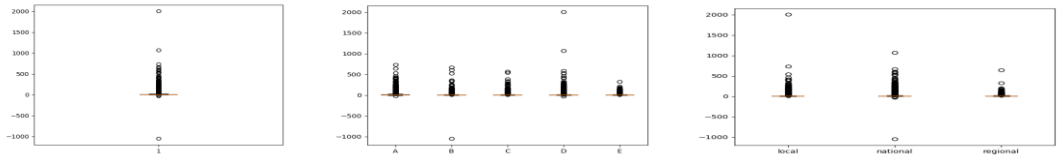


Fig 8: Transactions Boxplot - Overall and Categorical Breakdown (Type & Locale)

There are many outliers outside IQR*1.5 range for unit sales. The single negative value can be counted as a return. Additionally, the large observations could be attributed to many sources, but create the very large tail with outliers in the dataset. Compared to transactions where the outliers are only 4x the mean, unit sales mean is near 5 and outliers reach up to 2000 indicating outliers of 400x the mean value. Since there is larger variance within the number of sales within a single transaction this behavior can be attributed to a variety of transaction clusters. The locale and type distribution behavior look similar.

Time Series Analysis

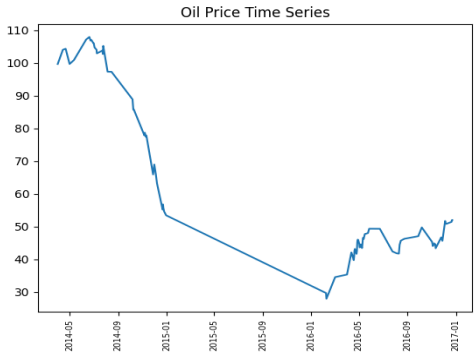


Fig 9: Time Series of Dcoilwtico

Oil prices have high variance, but volatility increases in later years. Additionally, there are breaks in the data where no data is collected. This accounts for seemingly non continuous data where the plot jumps. In 2014 the prices were very high, dropping the second half of 2014 and staying low through to the end of 2016. There is a large gap of data in 2015. We will not be analyzing the external geopolitical factors accounting for oil price fluctuations.

Determining the Distribution of Each Random Variable

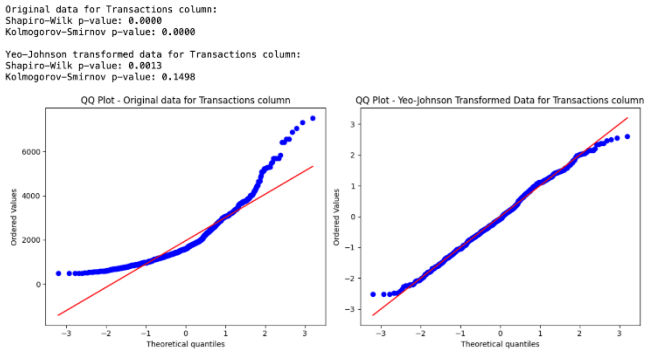


Fig 10: QQ-Plot for Transactions variable

We are going to use Yeo Johnson transformation in this section to make the data more normal, reduce the skewness in the data and addresses the issue of heteroscedasticity.

From the original data Transactions and Dcoilwtico are not normally distributed. After applying the Yeo-Johnson transformation, the "Transactions" data appears closer to normal distribution.

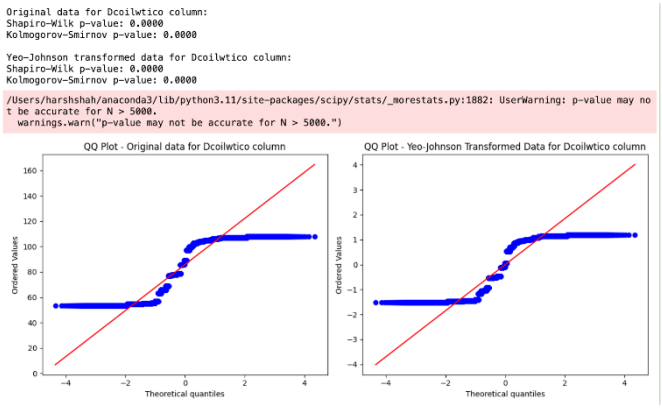


Fig 11: QQ-Plot for Dcoilwtico variable

The Q-Q plots show lack of normality, the original data's plot shows deviation from the red line which represents the expected distribution if the data is normal. The Yoe-Johnson transformed data still deviates from the line, which implies that even after transformation, the data does not follow a normal distribution.

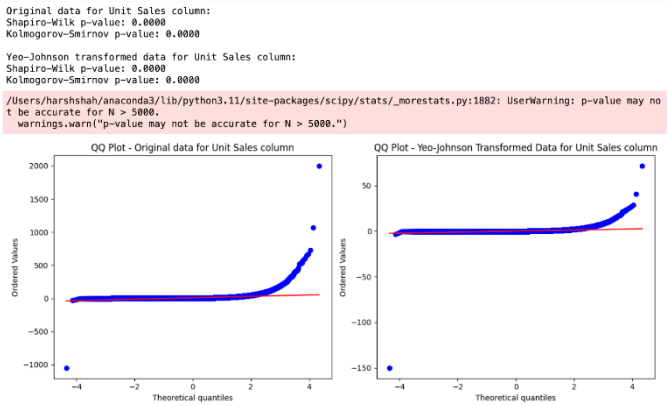


Fig 12: QQ-Plot for Unit Sales variable

Shapiro-Wilk and Kolmogorov-Smirnov indicate that both the original and the Yeo-Johnson transformed data do not follow a normal distribution, with p-values of 0.0000. statistical tests reject the hypothesis that the data is normally distributed.

Finally, we treated the outliers present using winsorization method, setting a threshold of 5^{th} and 95^{th} percentile.

Central Limit Theorem Exploration: Investigating Transactions

The Central Limit Theorem (CLT) states that for a sufficiently large sample size, the distribution of a sample variable approximates a normal distribution, and the sample mean and variance will be approximately equal to the mean of the whole population. Since $n \geq 30$ is generally considered to be sufficiently large, we study the Transactions column and consider one case in which the 1000-value column is divided into 10 sequential groups of n=100 samples each, and another in which it is divided into 10 simple random groups of n=100 samples each. Histograms and Q-Q plots were generated in each case, as well as a comparison between the population mean (μx) sample mean ($\mu_{\bar{x}}$), sample standard deviation ($\sigma_{\bar{x}}$), and population standard deviation ($\sigma_x/\sqrt{n}$).

Let us first investigate the random case.

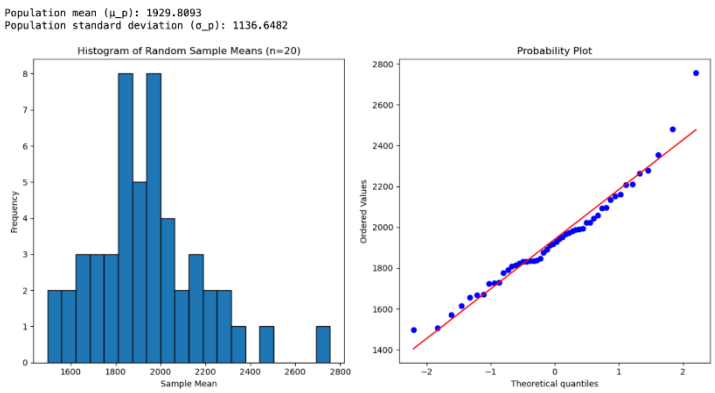


Fig 13: Histogram and QQ plot for Groups Divided Randomly (n=20)

The histogram shows the distribution if sample means calculated from sample of size 20. The population mean and the population standard deviation values help determine where the sample mean should center. The probability plot compares the sample means to a theoretical normal distribution. Points along the diagonal line indicates that it is a good fit to the normal distribution. The sample mean closely follows the line which implies that the distribution of sample mean is approximately normal, which confirms that under the conditions of Central limit theorem the sample mean is normally distributed around the population mean.

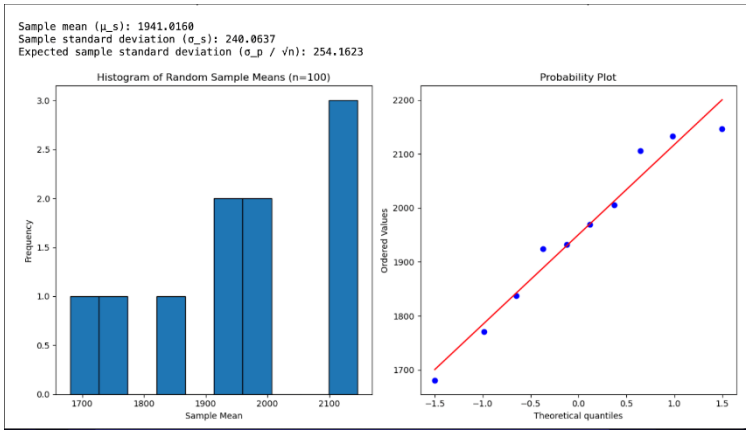


Fig 14: Histogram and QQ plot for Groups Divided Randomly (n=100)

The histogram here shows the distribution of sample mean collected from a sample of size 100. The data point aligns closely indicating the sample mean is normal. Both the histogram and the probability plot underscore the effect of Central Limit Theorem, meaning the distribution of the sample mean will approximate a normal distribution as the sample size increases.

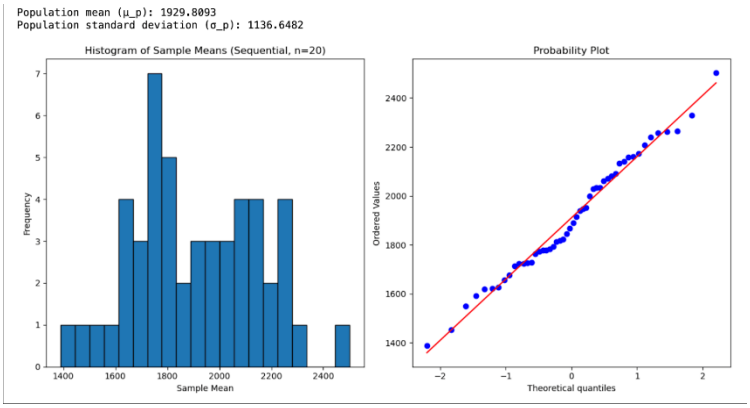


Fig 15: Histogram and QQ plot for Groups Divided Sequentially (n=20)

The histogram shows the distribution of sample mean with a sample size of 20. Sequential sampling might not cover the population distribution uniformly if the population has a spatial pattern. The sample mean is fairly aligned to the theoretical line which means that the distribution is normal. Variability in the sample mean when size is 20 might lead to a wider confidence interval which reflects a greater uncertainty in estimating the true population mean.

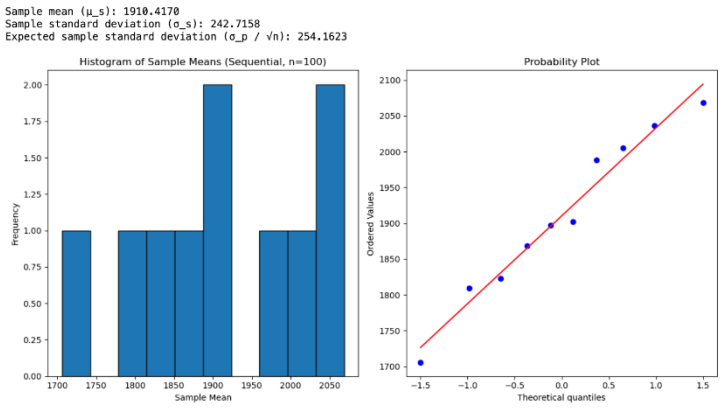


Fig 16: Histogram and QQ plot for Groups Divided Sequentially (n=100)

The histogram shows the distribution of sample mean with a sample size of 100. There is a variability in the sample mean which indicates different parts of the population may have different characteristics. The histogram and the probability plot support the confidence interval shown below. The proximity of the points in the probability plot indicates that the sample mean is normally distributed.

Constructing a Confidence Interval for Transactions from Random Samples

After selecting one of the random samples from the n=20 and n=100 cases, confidence intervals were constructed by using the critical t-value in the case of n=20 and using the Z-score approach for the n=100 case.

```
95% Confidence Interval for Transaction sample size 100: (1529.2739435183519, 1872.746056481648)
Sample population mean for Transaction sample size 100: 1701.01
True population mean for Transaction sample size 100: 1903.549
Sample Standard Deviation for Transaction sample size 100: 876.2204705610928
True Standard Deviation for Transaction sample size 100: 1058.587175629578

95% Confidence Interval for Transaction sample size 20: (1069.138077909557, 2381.961922090443)
Sample population mean for Transaction sample size 20: 1725.55
True population mean for Transaction sample size 100: 1903.549
Standard Deviation for Transaction sample size 20: 1497.7639289427773
True Standard Deviation for Transaction sample size 100: 1058.587175629578
```

Figure 17 – Confidence Intervals for n=20 and n=100 from Random Samples

The interval constructed from the larger sample (size 100) is expected to be more accurate (narrower) due to the decreased margin of error, demonstrating the influence of sample size on the precision of confidence intervals. This aligns with the Central Limit Theorem and principles of statistical inference, where larger sample sizes provide more precise estimates of population parameters.

Hypothesis Testing Using the Same Random Transactions Sample

Using the random sample extracted from the n=100 case as was for the confidence intervals calculated in the previous section, we perform the following two-tailed Z-test for a single sample and large sample size.

Test 1: For 10 sample with 100 values each
H0: $\boldsymbol{\mu} = 1900$
Ha: $\boldsymbol{\mu} \neq 1900$

Since the p-value (from Figure 18 for Test 1) is greater than 0.05, we accept the null hypothesis (H0), concluding that there is not sufficient statistical evidence that the sample mean differs from 1900 for the selected sample size. It does not necessarily mean that the true population mean is 1900, only that the sample data does not provide strong enough evidence to conclude a difference at the 0.05 significance level.

Test 2: For 50 sample with 20 values each
H0: $\boldsymbol{\mu} = 1900$
Ha: $\boldsymbol{\mu} \neq 1900$

Since the p-value (from Figure 18 for Test 2) is greater than 0.05, we accept the null hypothesis (H0), concluding that there is not sufficient statistical evidence that the sample mean differs from 1900 for the selected sample size. It does not necessarily mean that the

true population mean is 1900, only that the sample data does not provide strong enough evidence to conclude a difference at the 0.05 significance level.

Test 3: For 50 sample with 20 values each
H0: $\boldsymbol{\sigma} = 1100$
Ha: $\boldsymbol{\sigma} \neq 1100$

Since the p-value (from Figure 18 for Test 3) is greater than 0.05, we accept the null hypothesis (H0), concluding that there is not enough evidence to conclude that the population standard deviation is different from 1100. Similar to the mean, this does not prove that it is indeed 1100, but rather that the sample does not provide strong evidence against it.

Test 4: For 50 sample with 20 values each
H0: $\boldsymbol{\sigma} = 1100$
Ha: $\boldsymbol{\sigma} \neq 1100$
Since the p-value (from Figure 18 for Test 4) is greater than 0.05, we reject the null hypothesis (H0), concluding that there is not enough evidence to conclude that the population standard deviation is less than 1100. Similar to the mean, this does not prove that it is indeed 1100, but rather that the sample does not provide strong evidence against it.

```
Test 1: Sample size 100: t-statistic = 0.5566546216626699, p-value = 0.5790195976480433
Test 2: Sample size 20: t-statistic = 0.3257668960177601, p-value = 0.7481611868996239
Test 3: Chi-square statistic = 15.986706611570249, p-value = 1.3163295368537122
Test 4: Chi-square statistic = 15.986706611570249, p-value = 0.6581647684268561
```

Fig 18: Hypothesis Testing

Comparing Different Datasets: Type and Locale with Transactions

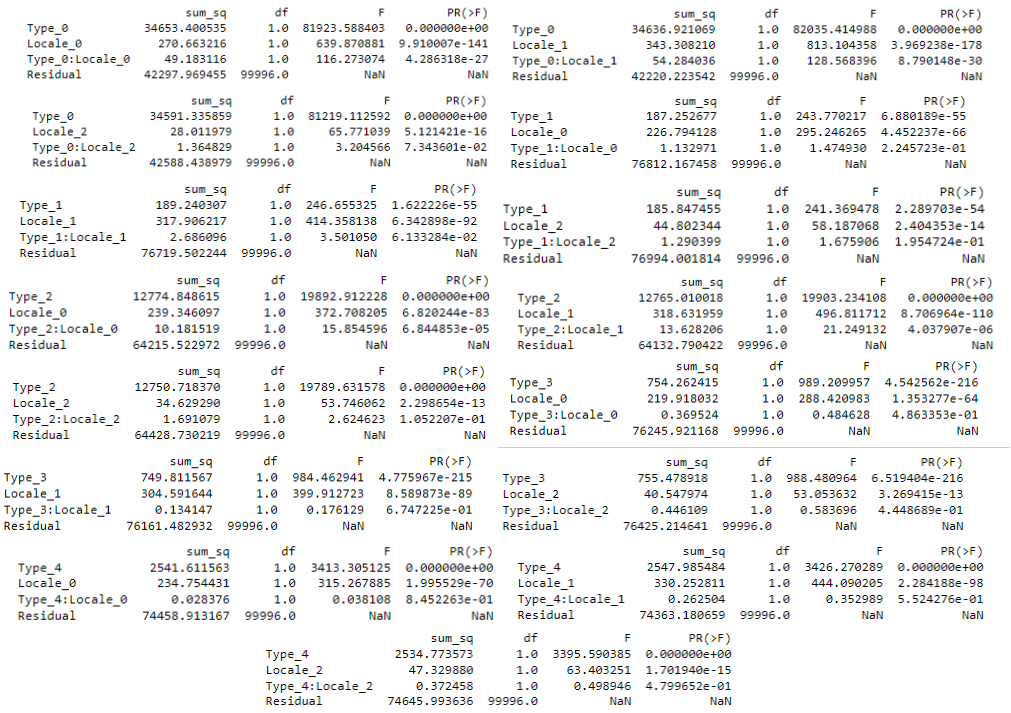
We have created a sample observation of only Type, Locale and Transactions random variable from the underlying dataset. We need to check if there is a significant effect of each of these variables on Transactions feature. Also, we need to check if the interaction of both variables is significant for the OLS model.

	sum_sq	df	F	PR(>F)
Type	21247.117933	1.0	38027.143772	0.000000e+00
Locale	117.973380	1.0	211.143493	8.653382e-48
Type:Locale	7.813008	1.0	13.983374	1.845372e-04
Residual	55871.322273	99996.0	NaN	NaN

Fig 19: ANOVA results for Type and Locale

From Figure X, we can deduct that the p-value for each of these variable and their interaction has significant effect on Transactions, as the p-values for all of them are less than our chosen significance level i.e. 0.05, suggesting using the interaction term while building the Linear Regression model.

Also, we need to check the significance of each variable after applying one hot encoding as it increases the dimension of the dataset. We need to check if any of the combination of Type and Locale is having any significant effect on Transactions or not.



	sum_sq	df	F	PR(>F)
Type_0	34653.400535	1.0	81923.588403	0.000000e+00
Locale_0	270.663216	1.0	639.870881	9.910007e-141
Type_0:Locale_0	49.183116	1.0	116.273074	4.286318e-27
Residual	42297.969455	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_0	34636.921069	1.0	82035.414988	0.000000e+00
Locale_1	343.308210	1.0	813.104358	3.969238e-178
Type_0:Locale_1	54.284036	1.0	128.568396	8.790148e-30
Residual	42220.223542	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_0	34591.335859	1.0	81219.112592	0.000000e+00
Locale_2	28.011979	1.0	65.771039	5.121421e-16
Type_0:Locale_2	1.364829	1.0	3.204566	7.343601e-02
Residual	42588.438979	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_1	187.252677	1.0	243.770217	6.880189e-55
Locale_0	226.794128	1.0	295.246265	4.452237e-66
Type_1:Locale_0	1.132971	1.0	1.474930	2.245723e-01
Residual	76812.167458	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_1	189.240307	1.0	246.655325	1.622226e-55
Locale_1	317.906217	1.0	414.358138	6.342898e-92
Type_1:Locale_1	2.686096	1.0	3.501050	6.133284e-02
Residual	76719.502244	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_1	185.847455	1.0	241.369478	2.289703e-54
Locale_2	44.802344	1.0	58.187068	2.404353e-14
Type_1:Locale_2	1.290399	1.0	1.675906	1.954724e-01
Residual	76994.001814	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_2	12774.848615	1.0	19892.912228	0.000000e+00
Locale_0	239.346097	1.0	372.708205	6.820244e-83
Type_2:Locale_0	10.181519	1.0	15.854596	6.844853e-05
Residual	64215.522972	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_2	12765.010018	1.0	19903.234108	0.000000e+00
Locale_1	318.631959	1.0	496.811712	8.706964e-110
Type_2:Locale_1	13.628206	1.0	21.249132	4.037907e-06
Residual	64132.790422	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_2	12750.718370	1.0	19789.631578	0.000000e+00
Locale_2	34.629290	1.0	53.746062	2.298654e-13
Type_2:Locale_2	1.691079	1.0	2.624623	1.052207e-01
Residual	64428.730219	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_3	754.262415	1.0	989.209957	4.542562e-216
Locale_0	219.918032	1.0	288.420983	1.353277e-64
Type_3:Locale_0	0.369524	1.0	0.484628	4.863353e-01
Residual	76245.921168	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_3	749.811567	1.0	984.462941	4.775967e-215
Locale_1	304.591644	1.0	399.912723	8.589873e-89
Type_3:Locale_1	0.134147	1.0	0.176129	6.747225e-01
Residual	76161.482932	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_3	755.478918	1.0	988.480964	6.519404e-216
Locale_2	40.547974	1.0	53.053632	3.269415e-13
Type_3:Locale_2	0.446109	1.0	0.583696	4.448689e-01
Residual	76425.214641	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_4	2541.611563	1.0	3413.305125	0.000000e+00
Locale_0	234.754431	1.0	315.267885	1.995529e-70
Type_4:Locale_0	0.028376	1.0	0.038108	8.452263e-01
Residual	74458.913167	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_4	2547.985484	1.0	3426.270289	0.000000e+00
Locale_1	330.252811	1.0	444.090205	2.284188e-98
Type_4:Locale_1	0.262504	1.0	0.352989	5.524276e-01
Residual	74363.180659	99996.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Type_4	2534.773573	1.0	3395.590385	0.000000e+00
Locale_2	47.329880	1.0	63.403251	1.701940e-15
Type_4:Locale_2	0.372458	1.0	0.498946	4.799652e-01
Residual	74645.993636	99996.0	NaN	NaN

Fig 20: ANOVA results for Type and Locale after using One Hot Encoding

From Figure X, we were able to understand the different interactions which are significant for the model training. We observed that interaction between {Type_0, Locale_0}, {Type_0, Locale_1}, {Type_2, Locale_0} and {Type_2, Locale_1} have significant effect on Transactions. It suggests that the relationship which we found previously that the interaction between Type and Locale has significant effect on Transactions was due to the interaction between (0,0), (0,1), (2,0) and (2,1) pairs only.

Model Implementation and Results

We are predicting Transactions Variable using Linear Regression. As we know, linear regression is sensitive to outliers, multicollinearity, and the variance in the dataset. Hence, before applying Linear Regression we have already treated outliers in the dataset, multicollinearity and have transformed data using Standard Scaler to reduce the variance in the dataset. Also, from the above two-way ANOVA analysis, we are using the interaction term in our model too.

OLS Regression Results

Dep. Variable:	Transactions	R-squared:	0.315
Model:	OLS	Adj. R-squared:	0.315
Method:	Least Squares	F-statistic:	9178.
Date:	Tue, 30 Apr 2024	Prob (F-statistic):	0.00
Time:	19:02:26	Log-Likelihood:	-1.1009e+05
No. Observations:	100000	AIC:	2.202e+05
Df Residuals:	99994	BIC:	2.202e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6416	0.007	97.501	0.000	0.629	0.654
Type	-0.3325	0.003	-117.139	0.000	-0.338	-0.327
Locale	0.0742	0.007	10.041	0.000	0.060	0.089
Type:Locale	-0.0122	0.003	-3.863	0.000	-0.018	-0.006
Dcoilwtico	-0.0100	0.003	-3.985	0.000	-0.015	-0.005
unit_sales	0.5671	0.008	74.482	0.000	0.552	0.582

Omnibus:	5504.718	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6462.166
Skew:	0.620	Prob(JB):	0.00
Kurtosis:	3.116	Cond. No.	13.5

Fig 21: OLS Regression Results

From the Figure, we can observe that the model has good F-statistic value and p-value 0 for all the features present but it's R2 is very small (0.315) i.e. the model is not able to explain the variance in the Transactions variable.

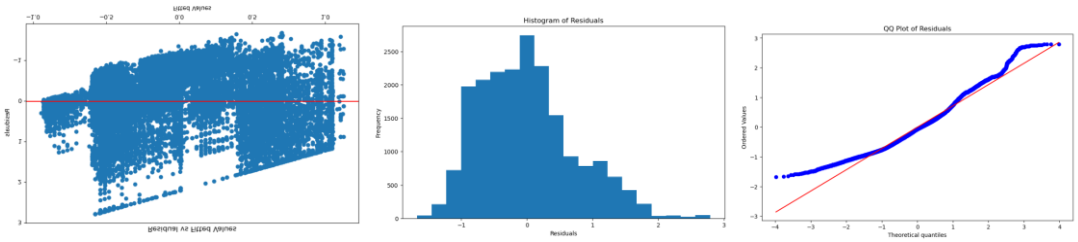


Fig 22: Residual VS Fitted Data, Histogram of residuals, QQ-Plot of predicted values using OLS Regression

We can observe from the above charts that residuals show some pattern, not entirely random. Also, in QQ plot, points deviate from the line, suggesting non-normality in the residuals i.e. this model is not able to predict the Target variable with the given features and needs further investigation.

Now applying Polynomial to check if it has non-linear relationship. If it has no-linear relationship, we can expect an increase in the R2 value and the QQ plot will be much closer to the straight line. We used GridSearchCV for obtaining the best hyper parameters for this model and obtained it at n=3 with an R2 value of 0.59.

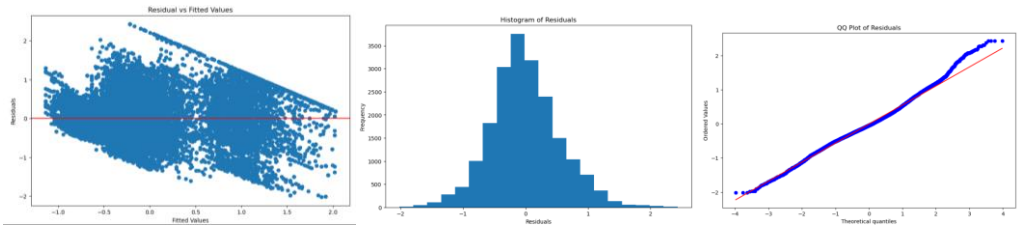


Fig 23: Residual VS Fitted Data, Histogram of residuals, QQ-Plot of predicted values using Polynomial Regression (n=3 HPO)

We can clearly observe that residuals appear randomly scattered over the axis. Also, the QQ plot is much closer to the straight line suggesting the normal distribution of residuals. Hence, we can note that with the underlying features the target variable, instead has a non-linear relationship.
Now applying DT regressor, without applying outlier treatment as it doesn't affect DT. We used GridSearchCV for obtaining the best hyper parameters for this model and obtained it at max_depth=9, min_samples_split=5 with an R2 value of 0.69.

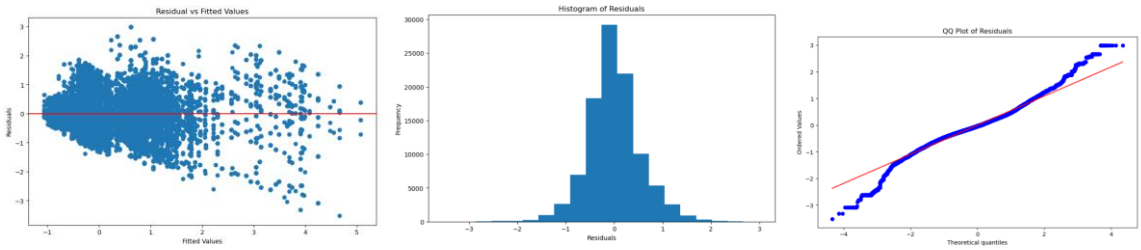


Fig 24: Residual VS Fitted Data, Histogram of residuals, QQ-Plot of predicted values using Decision Tree Regressor (HPO)

Clearly, from these plots we can easily conclude that Polynomial is able to undersand and predict the model more accurately than DT Regressor. The QQ-plot are deviating from the line at the tails. Even though the model can explain more variance in the Target variable (0.69) than Polynomial Regression (0.59), residual plot shows more scatter in polynomial and more stable QQ-plot. Hence, we can conclude that Polynomial Regression is able to understand the features more accurately than these other models.

Now we use One Hot Encoding and apply everything again to check the improvements in the model. As the qualitative variables consists of more than 2 unique values, we are now using OHE (One Hot Encoding) which will increase the total number of features. From the two-way ANOVA analysis, we are also considering the interaction of all the newly created features if there are any.

OLS Regression Results

Dep. Variable:	Transactions	R-squared:	0.545
Model:	OLS	Adj. R-squared:	0.545
Method:	Least Squares	F-statistic:	9999.
Date:	Tue, 30 Apr 2024	Prob (F-statistic):	0.00
Time:	19:04:28	Log-Likelihood:	-89549.
No. Observations:	100000	AIC:	1.791e+05
Df Residuals:	99987	BIC:	1.792e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0637	0.004	-18.140	0.000	-0.071	-0.057
Type_0	1.1198	0.016	69.738	0.000	1.088	1.151
Type_1	-0.0339	0.007	-4.707	0.000	-0.048	-0.020
Type_2	-0.5898	0.016	-37.753	0.000	-0.620	-0.559
Type_3	-0.0354	0.006	-5.524	0.000	-0.048	-0.023
Type_4	-0.5244	0.008	-62.355	0.000	-0.541	-0.508
Locale_0	-0.0510	0.005	-10.467	0.000	-0.061	-0.041
Locale_1	0.0485	0.005	10.545	0.000	0.039	0.058
Locale_2	-0.0612	0.007	-8.818	0.000	-0.075	-0.048
Type_0:Locale_0	-0.0260	0.022	-1.196	0.232	-0.069	0.017
Type_0:Locale_1	0.0786	0.021	3.704	0.000	0.037	0.120
Type_2:Locale_0	-0.0013	0.021	-0.061	0.952	-0.043	0.040
Type_2:Locale_1	-0.0402	0.021	-1.938	0.053	-0.081	0.000
Dcoilwtico	-0.0007	0.002	-0.351	0.726	-0.005	0.003
unit_sales	0.3846	0.006	61.450	0.000	0.372	0.397

Omnibus:	5382.300	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7593.831
Skew:	0.496	Prob(JB):	0.00
Kurtosis:	3.916	Cond. No.	2.40e+15

Fig 25: OLS Regression Results (OHE)

Clearly, using OHE, creating more dimensions, the OLS model can understand the data more thoroughly by increasing the R2 value to 0.545 and more F-statistic value for the model. But we can also observe that by introducing OHE, the Dcoilwtico, Type_0: Locale_0 and Type_2: Locale_0 features are not significant for this model with high p-values more than our chosen significance level i.e. 0.05.

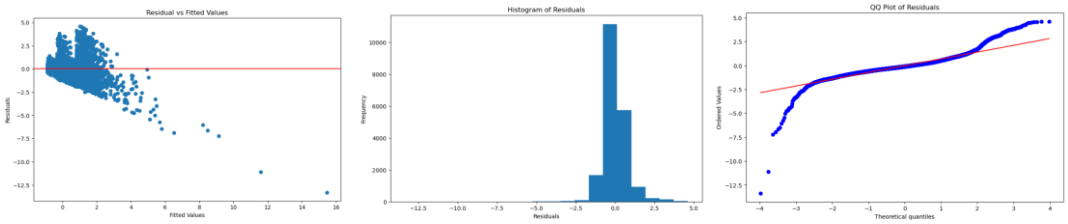


Fig 26: Residual VS Fitted Data, Histogram of residuals, QQ-Plot of predicted values using OLS Regression (OHE)

The residual plot confirms that OLS models are sensitive to dimensions. Even the QQ plot seems to be deviating a lot from the straight line. OHE is not able to improve the performance of the model.
Let's see now how OHE is going to affect Polynomial Regression as it is sensitive to dimensions.

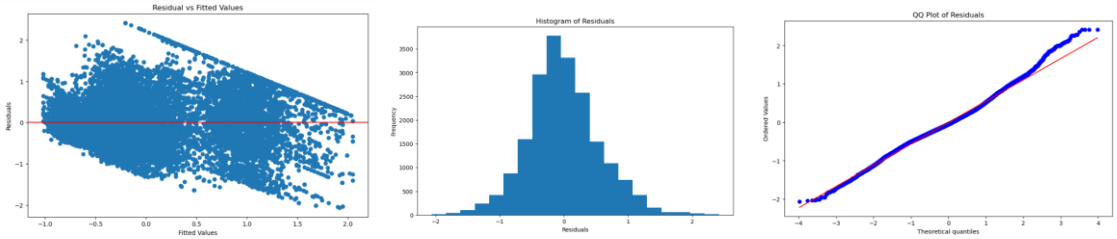


Fig 27: Residual VS Fitted Data, Histogram of residuals, QQ-Plot of predicted values using Polynomial (n=3 HPO OHE)

The R2 value improved to 0.60 and the residual plot seems to be more scattered than the previous one, suggesting that introducing OHE increased the learning of Polynomial model Now again applying DT Regressor with HPO.

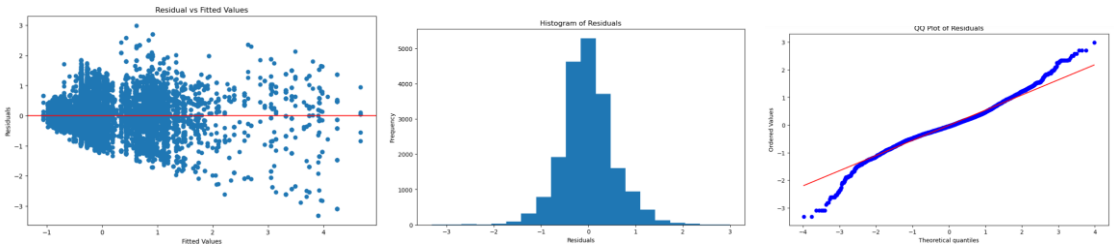


Fig 28: Residual VS Fitted Data, Histogram of residuals, QQ-Plot of predicted values using Decision Tree Regressor (HPO OHE)

Clearly, increasing dimensions is not an issue for DT Regressor and we can see from the residual plot that they are more scattered than the previous model. But Polynomial regression is outperforming the DT here.

Conclusion

Upon investigation we found out that before applying the model we had to treat the data accordingly. Outliers are treated using Yeo-Johnson method, Qualitative was handled by using OHE and as there was no highly collinear pairs, we were not needed to handle multi collinearity. Histograms were used to visualize the distribution of the variables and confirmed through QQ plots and KS tests.
Also, we found out that after using OHE and the number of dimensions is increased, we needed to apply two way ANOVA for all pairs possible and check the significance of each pair and their interaction before we apply the model as relevant interaction could help learn the model better which we observed in the previous section. Also, we realized that Transactions has a nonlinear relationship with its predictors, leading to a much better result with Polynomial Regression.
Decision Tree Regressor, even after increasing the number of dimensions was not able to learn from these predictors better than Polynomial which was clearly visible in the scattered residual plot. Even though DT can explain 69% of the variance of the target variable than 60% of the polynomial regression but the spread in residual plot suggests otherwise.