MGCR-271-001 Business Statistics

Regression Analysis on the Factors Affecting the Monthly Spending of Desautels Student

Kaibo Zhang: 261110409

Desautels Faculty of Management

McGill University

April 13th, 2023

Introduction

University presents an opportune platform for students to develop their financial literacy skills. At Desautels, the Faculty of Commerce, students begin their education by learning about the economy, accounting, and global business backgrounds, which are on broader scopes than trifles like monthly spending. Despite most of their outstanding performance academically, many encounter challenges in controlling their expenditures due to an inadequate understanding of the underlying factors affecting their spending behavior. Thus, this research investigates the effects of multiple factors on the monthly spending of students enrolled at Desautels. By performing statistical analyses of the data collected, this paper aims to interpret the relationship between our targeting variables and monthly spending to provide valuable insights into practical recommendations for controlling students' monthly spending. These recommendations will assist students in understanding their spending behaviors and managing their budgets to attain their respective financial goals.

Data Collection and Justifications for the Appropriateness of Sampling

Our methodology involves an online survey utilizing a self-administered questionnaire developed on the Qualtrics platform.

The questionnaire comprised seven questions, six requiring numerical inputs, while the remaining required a text input. Based on conventional wisdom, we selected the following independent variable: the distance of residence from McGill in kilometers (quantitative), frequency of dining out per month in numbers of times (quantitative), living arrangement (categorical), number of monthly subscriptions (quantitative), frequency of going to the groceries in numbers of times (quantitative), and the number of shopping (quantitative). The

following text will refer to them as dine-out, subscription, groceries, and shopping, respectively. The survey was distributed using both social media by sharing the links of the questionnaire and the traditional in-person method.

In the questionnaire, we strategically introduced the survey's purpose, scope, and confidentiality policy to the respondents before the actual survey contents so they could clearly understand the purpose and the importance of their faithful response to the results. Secondly, we kept the wording of the questions simple, direct, and free from ambiguity to avoid misleading or confusing our respondents. This ensures the respondents' answers align with the situation and protects our data collection from certain biases.

For our method, we faced the following biases. Firstly, such a collection method does not guarantee a simple random sample. The undercoverage bias may influence our results as we did not perform random sampling that ensures every representative has an equal chance to be selected in our sample. We may include more people in the same group and neglect some individuals in other categories. Secondly, rather than randomly sending our questionnaire, we partially shared the questionnaire with our friends. Their consumption behavior may overly influence our results and demonstrate patterns caused by lurking variables not included in our research.

Although we could not eliminate the biases that occurred, we still managed to obtain a large enough sample size where we could apply the Central Limit Theorem and the rule of thumb to ascertain that our sample could be a valid representation of the entire population. At the same time, our questionnaires are anonymous to get more faithful responses.

Describing the Data

1. Monthly spending

a. Describing the Distribution

Our results indicate that the distribution of monthly spending is positively skewed. 75% of the data points are clustered between $170 and $2,200, with the remaining widely dispersed in the ranges between $2,200 and $5,000. This skewness pulls the sample mean to the left-hand side of its median, affecting the shape of the distribution. This is anticipated as the prices and price-related data distributions are typically positively skewed. The IQR Rule also identifies two extreme outliers with monthly spending values of $5,000 and $4,300, respectively (see Appendix A).

b. Inference for Population Mean, μ

The mean and standard deviation are known for our sample. A sample size of n=99 also passes the threshold of n>25 for applying the Central Limit Theorem. All the preconditions for a one-sample z-test for population mean are met. With these assumptions in mind, we assume that the distribution of the monthly spending for students at Desautels is approximated normally distributed as N~($1,645.66, $1,063.63). With the calculated confidence interval, we are 95% confident that the true population mean μ lies between the range of [$1,436.14, $1,858.18].

c. Comparison of the Monthly Spending Distribution on Residence Type

Both distributions appear skewed to the right since their means exceed the medians. However, the skewness of the distribution of monthly spending on rent appears to be more significant. Each distribution has a higher extreme outlier above Q3. Students who rent have a higher mean in their monthly spending habits than those who live in dorms, and vice versa for

their medians. In addition, more spread and variations are observed in the distribution of those who rent apartments or condos, as seen by its larger variance and standard deviation (see Appendix B). This pattern is likely caused by the differences in the ranges of the two distributions. Among the students who rent, their monthly spending has a more extensive range than those students living in dormitories ($4,830 > $3,400). Similarly, zooming into the interquartile range, 50% of the samples who live in the dorm fall between $2,000 and $1,000, while 50% of the dataset who rents fall between $869 and $2,325.

Noticeably, the range of monthly spending for students living in dormitories falls within the range of monthly spending for students who rent (see Appendix C). This observed phenomenon is most likely due to the greater variability of rental options compared to the limited availability of dormitories near the campus of McGill University. Dorms are typically clustered around McGill's campus. This prominently confines students' consumption choices, resulting in similar spending patterns. Conversely, rental listings are dispersed around various neighborhoods in Montreal, thus allowing for a wider breadth of options for consumption. In summary, students who rent may have access to a broader range of consumption options, such as grocery stores, restaurants, or shopping centers, that are not easily accessible to students living in dormitories due to their location or limited mobility. This may contribute to the observed phenomenon in monthly spending patterns between these two groups of students.

2. Correlation between Variables:

This section relies on the calculated results using the CORREL function in EXCEL for studying and interpreting the correlation coefficients (r) between independent variables.

Based on our analysis, the independent variables we examined are very weakly correlated, as evidenced by the fact that most correlation coefficients fall into the range of [-0.2, 0.2]. This suggests that there is little to no relationship between these variables, indicating that potential associations between the independent variables will not cast a significant impact on multiple regression and, thus, will not shadow our results (see Appendix D).

Though their associations are weak, patterns that coincide with conventional expectations are still present in our dataset. Individuals' distance to McGill tends to be negatively correlated with Grocery, Dining out, Shopping, and Subscriptions. This suggests that increased students' distance to McGill will likely trigger a corresponding decrease in the variables mentioned above. This is unsurprising given that McGill is in the downtown area, giving the students living here more exposure to consumption activities, while students residing farther from McGill may have fewer amenities and opportunities for consumption. Apart from this pattern between certain variables, all others are positively correlated, moving in the same direction in response to a change in another.

In short, the correlation coefficients (r) have shown weak or almost no correlations between the independent variables included in the research. While some positive and negative patterns can be observed, it is essential to note that these correlations will not mislead and shatter the ultimate analysis of the independent variables' power in predicting the dependent variable.

Hypothesis Testing

1. Mean Difference in Monthly Spending on Residence Type

Residence type is always considered to be an influential factor in determining individuals' monthly spending habits. Different residence types can be associated with differences in access to amenities. Students who rent may have different access to amenities like parks, restaurants, shopping centers, and entertainment venues, which in turn, can impact their expenditure on all kinds of activities. To arrive at a statistically significant conclusion about the validity of this statement, we conducted a two-sided t-test to make inferences about the true differences in population means of our interested subject.

With $n_1 = 49 > 25$ and $n_2 = 50 > 25$, we assume that the sample σ is approximately equal to the population σ , meeting all specified requirements for performing a t-test at a significance level of $\alpha = 0.05$.

Let μ_1 = \$1701.63 denote the mean monthly spending of students who are renting.

Let μ_2 = \$1588.55 denote the mean monthly spending of students who live in a dorm.

$H_0: \mu_1 - \mu_2 = 0$ Ha: $\mu_1 - \mu_2 \neq 0$

A t-statistic of 0.5285 at a degree of freedom of 48 ($df = n_1 - 1$) corresponds to a p-value greater than 0.5. Assuming that the null hypothesis is true, the probability of seeing data as extreme or more extreme than what was observed is far-fetched from our significance level. Therefore, we fail to reject the null hypothesis in favor of the alternative hypothesis and conclude that, based on our data, students with different residence types do not significantly differ in their monthly spending. One possible reason that justifies this result is that contrary to its effect on those who work, residence type may not strongly influence students at Desautels since most of their budget constraints are subjected to financial aid from their parents, which does not have substantial variation in between. Moreover, although students

living in dorms may spend less because of mean plans, students who rent may have access to cheaper off-campus housing options, while students who live in a dorm may be required to pay a fixed rate for their housing. These factors could balance out and lead to similar monthly spending habits between the two groups.

2. Relationship between Distance and Monthly Spending

It is commonly assumed that distance and monthly spending are positively related, as living closer to McGill reduces transportation costs and more opportunities for comparison shopping because McGill is in the downtown area. However, the result from running a linear regression presents a beta contrary to our expectations. Therefore, we are interested in testing the validity of this result using a hypothesis testing for the slope of the line of best fit that is plotted between distance and monthly spending. Our hypothesis will test whether the linear relationship is significant at a significance level of 0.05. The residual plot demonstrates a roughly random pattern for the residuals, proving the appropriateness of using a linear model.

$$H_0: \beta_1 = 0 \qquad Ha: \beta_1 \neq 0$$

At a degree of freedom of 97 (df=n-2), the recorded p-value for a one-sided test is 0.0731, which is slightly greater than α (see Appendix E). Therefore, we conclude that there is not enough strong evidence of a linear relationship between distance to McGill and students' monthly spending. However, given the p-value would pass at a 10% significance level, we suspect that a negative linear relationship may still exist in our sample. This negative slope may be attributed to the higher costs of living in downtown areas. The cost of daily activities tends to be higher in downtown areas. Students living near McGill may have to bear these higher costs, which could explain this result.

3. Appropriateness of Establishing Multiple Regression Model

Since one of the primary purposes of this paper is to establish the best model for predicting our dependent variable, we want to know if the independent variables contained in this study have an effect on the monthly spending of students at Desautels. Therefore, we want to test the following hypothesis at a significance level of $\alpha = 0.05$:

H_0: All independent variables have no effects on monthly spending

Ha: At least one independent variable leads to different monthly spending

Running a global F-test, the multiple regression that takes in all independent variables arrives at a p-value of 0.002695, providing significant evidence to reject the null hypothesis in favor of the alternative hypothesis (see Appendix F). Therefore, we conclude that at least one of the independent variables has an influence on monthly spending. Thus, it is appropriate and meaningful to proceed with establishing the best multiple regression model.

The Best Multiple Regression Model

This research conducted six separate simple linear regression analyses to identify the optimal combination of independent variables that provides the most predictive power over the dependent variable in our study. Each analysis examined the relationship between one independent variable and the monthly spending of Desautels students, measured in units of the Canadian dollar. By utilizing this approach, we aimed to determine the most efficient combination of independent variables to be incorporated into a multiple regression model. Following the standard conventions and to simplify the analysis, we omitted the variable (Rent) with a p-value exceeding 0.2 (0.5994>0.2). We ranked the remaining variables based on their influence size, as determined by their R^2 values (from largest to smallest) and

corresponding p-values (from smallest to largest). Subsequently, we began from the highest-ranking variable to run multiple regression and examined the related change in the Adjusted R^2. We continued this process for each subsequent variable, in descending order of influence ranking, to determine whether it should be included in the final model (see Appendix G). The adjusted R^2 increased sequentially when adding groceries, subscriptions, and dine-out, resulting in 0.158616. However, adding shopping as the fourth variable decreased adjusted R^2 by 0.005372 (0.158616- 0.153244). Two further regression analyses were conducted, with one adding a fifth variable resulting in a minor improvement to 0.159447, and the other adding distance as a fourth variable in the three-variable model, leading to an adjusted R^2 of 0.16488. These results suggest that adding shopping to the model slightly diminished its predictive ability. While including distance as a variable proved to be a more substantial predictor of the dependent variable than distance, adding a fourth variable to the model results in an increase in overall predictive power that is negligible. Thus, the multiple regression that takes in groceries, the number of subscriptions, and dine-out are determined to be the best model for predicting the monthly spending of Desautels students. The equation of the regression is provided as follows, with variables all measured in numbers of (see Appendix H):

Monthly Spending = 916.4253 + 16.7489 (Dine-out) + 85.4177 (Groceries) + 29.4629 (Subscriptions) + *e*

This can be interpreted that for the students at Desautels,

b_1: adding one time of dine-out increases their expected monthly spendin$_2$g by $916.4253 (holding the other two factors constant)

b_2: adding one time of going to the groceries increases their expected monthly spending by $85.4177 (holding the other two factors constant)

b_3: every additional channel they subscribed to increases their expected monthly spending by $29.4629 (holding the other two factors constant)

Recommended Solution

According to our research, Desautels' students' monthly spending is mainly related to the frequency of dining out, the number of subscriptions, and groceries. We would provide recommendations to Desautels' students to help them effectively manage their finances by controlling each spending area.

Our research indicated that a higher frequency of grocery shopping could drive a higher monthly expense. Purchasing a meal plan is mandatory for students residing in McGill's dormitories. Consequently, they are expected to go to the groceries less often than those living in rental apartments, who have the option of cooking their meals. However, a p-value greater than 0.5 for a double-sided hypothesis testing for differences in mean infers that students of different residence types do not have different frequencies of grocery shopping (see Appendix I). This brings speculation that students' primary incentive for grocery shopping: they go to groceries for snacks or beverages rather than essential food items like fruits and vegetables. Therefore, we recommend that students manage the number of times they go to the grocery store appropriately, saving their budgets and developing healthier eating habits.

In addition, our research indicates that students who frequently dine out tend to have higher monthly expenditures due to the higher average cost of restaurant meals than those

offered by the university's dining halls. As a potential solution, we recommend that students consider purchasing meal plans, which can be a cost-effective way to reduce spending. The dining hall at McGill University could work on creating new dishes and resetting the prices in a more affordable way to attract more potential students to consume.

Moreover, our survey findings indicate that the number of subscriptions contributes to students' monthly expenses. To alleviate financial pressure on students, the faculty could consider offering academic-related subscriptions for free, such as Grammarly or Microsoft Office. This approach would not only ease students' financial burden but also contribute to their academic success, which is essential to the mission of educational institutions.

Areas of Improvement

Several concerns can be addressed to refine the research further to obtain a more comprehensive and accurate prediction model with reliable results.

1. Incorporating More Independent Factors

Our established multiple regression model can only explain 18.44% of the variation in Desautels students' monthly spending. This suggests that other outstanding variables may have a more potent influence on our dependent variable. Other factors such as income, ethnicity, cultural background, and gender could also be influential. Therefore, including potential lurking variables, either based on demographic factors or theoretical frameworks, may give higher predictive power to our regression model.

2. Enlarging Sample Size for Improved Diversity and Comprehensiveness

As mentioned earlier, our data collection methods are currently limited and subject to various biases. Furthermore, our sample fails to reach the local Montreal population who live

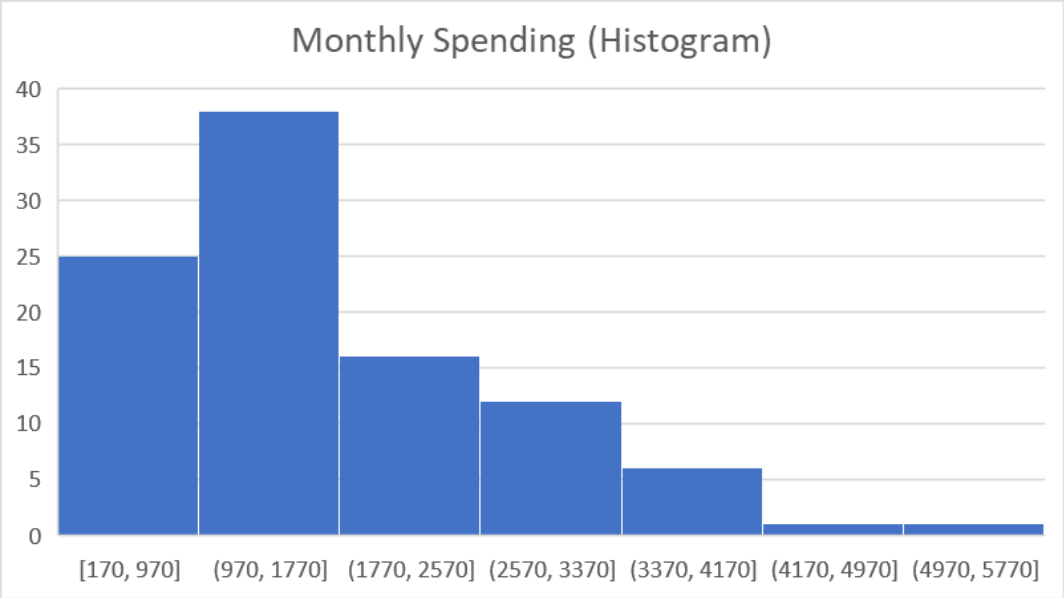
in their own home. This exclusion can lead to inaccurate conclusions and a lack of representation for this group. To address this problem, our sample size can be enlarged to reach a more diverse and representative sample. More specifically, we could collaborate with club organizations such as SSMU to reach these underrepresented populations in our data collection. By taking a more inclusive and diverse approach to data collection, we can improve the accuracy and reliability of our findings through a sample that is more representative of its population.

Conclusion

Though the ultimate model does not provide appealing accounts for the variations in the dependent variable under investigation, patterns exhibited still provide insights in conjunction with people's general instincts. As the betas of the final model have suggested, monthly spending tends to move in the same direction with positive changes in dine-out, groceries, and subscriptions. Controlling and reducing the numbers related to these three variables can control spending to a certain extent. At the same time, the comparably small R^2 also implies that the conventional wisdom we commonly adhere to may not hold as much impact as we once believed. This finding highlights the importance of conducting academic research and referring to pre-established theories, if necessary, to identify the key drivers of consumer behavior and avoid relying solely on preconceived notions or anecdotal evidence. Other research on similar topics tests the significance of demographic factors and theoretical models and is able to explain more of the variation in monthly spending. As it stands today, the breadth and depth of studies can be extended to analyze other variables that may have significant effects on the financial habits of our target population.

Appendix A:

The Histogram and Boxplot of the Monthly Spending Distribution



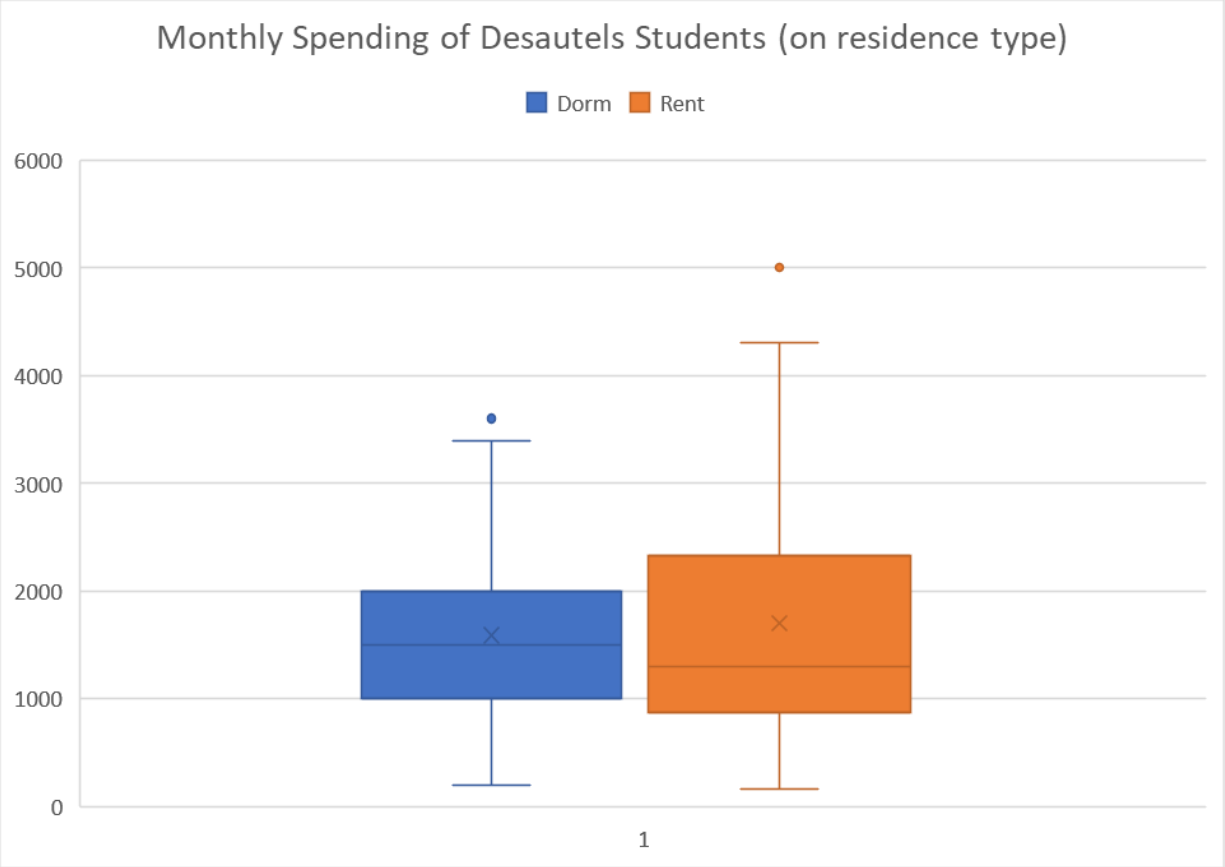
Appendix B

Table of the Number Summary of Monthly Spending on Residence Type

Number Summary of Monthly Spending (on residence type)								
	Mean	Variance	Sample std	Min	Q1	Q2(Median)	Q3	Max
Rent	1701.63	1459484.7435	1208.0914	170.00	869	1300	2325	5000.00
Dorm	1588.55	813286.6276	901.82406	200.00	1000	1500	2000	3600.00

Appendix C

Boxplot Comparison for Monthly Spending on Residence Type



Appendix D:

Table of Correlation Coefficients Between Independent Variables

Correlation between variables					
	Distance	Dine out	#Subsriptions	Groceries	Shopping
Distance	1				
Dine out	-0.0828804	1			
#Subsriptions	-0.0535803	0.13176107	1		
Groceries	-0.1269376	0.12125916	0.03831567	1	
Shopping	-0.0532697	0.18347229	0.18415087	0.2481267	1

Appendix E:

Excel Output of Linear Regression on Distance against Monthly Spending

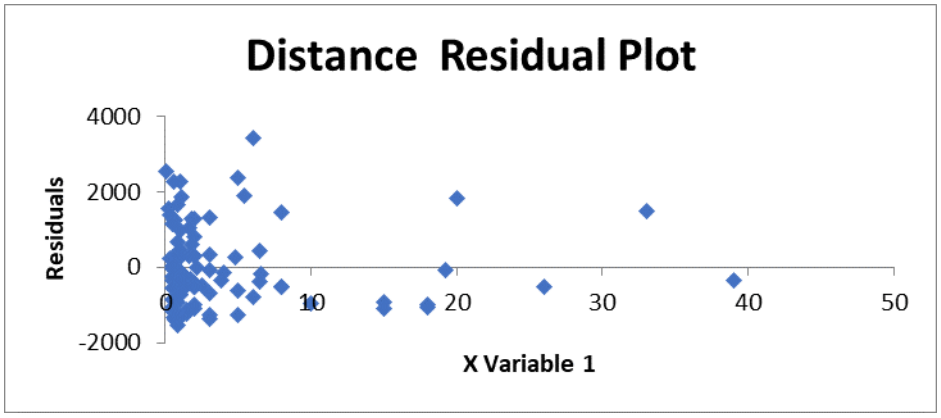
SUMMARY OUTPUT

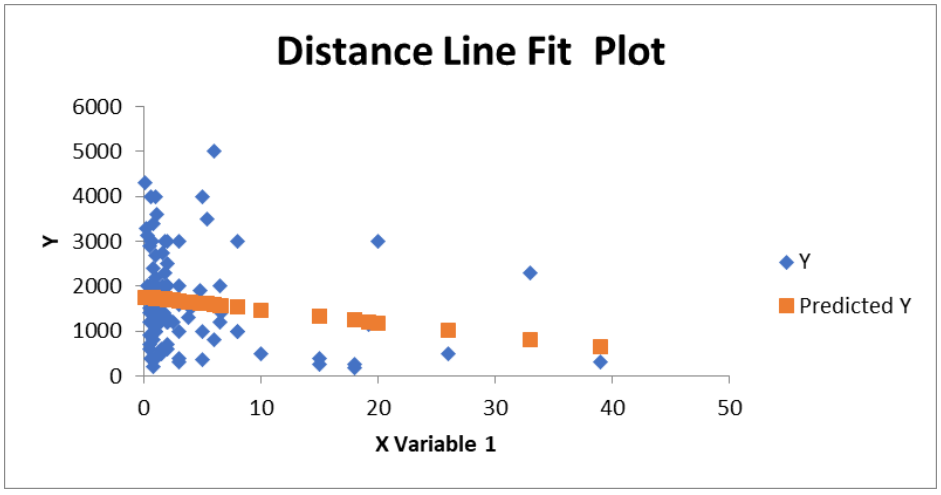
Regression Statistics	
Multiple R	0.180872574
R Square	0.032714888
Adjusted R Square	0.022742876
Standard Error	1051.468715
Observations	99

ANOVA

	df	SS	MS	F	Significance F
Regression	1	3627065.333	3627065.333	3.280671	0.073195266
Residual	97	107241886.5	1105586.459		
Total	98	110868951.8			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1759.877595	123.0607823	14.30088093	1.26E-25	1515.636007	2004.119183	1515.636007	2004.119183
Distance	-28.72885964	15.86123697	-1.811262242	0.073195	-60.20902441	2.751305138	-60.20902441	2.751305138





Appendix F:

Excel Output of Global F-test

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.438537087				
R Square	0.192314777				
Adjusted R Square	0.139639653				
Standard Error	986.5795883				
Observations	99				
ANOVA					
	df	SS	MS	F	Significance F
Regression	6	21321737.7	3553623	3.650960161	0.002695288
Residual	92	89547214.12	973339.3		
Total	98	110868951.8			

Appendix G:

Tables of Hierarchy Ordering and Summary of R^2 and P-value

Hierarchy Ordering	
	Rank
Distance	5
Dine out	3
Rent	6
#Subsriptions	2
Groceries	1
Shopping	4

Table Summary of R^2 and P-value						
α=0.05	Distance	Dine out	Rent	#Subsriptions	Groceries	Shopping
R Square	0.0327	0.0602	0.0029	0.0680	0.0897	0.0389
P-value	0.0732	0.0144	0.5994	0.0091	0.0026	0.0504

Appendix H:

Related Outputs of the Best Multiple Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.429386448
R Square	0.184372721
Adjusted R Square	0.15861607
Standard Error	975.638746
Observations	99

ANOVA

	df	SS	MS	F	Significance F
Regression	3	20441210.36	6813737	7.158257	0.000219546
Residual	95	90427741.46	951871		
Total	98	110868951.8			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	916.4252756	189.3933998	4.838739	5.04E-06	540.4318385	1292.418713	540.4318385	1292.418713
Dine out	16.74892741	8.620022263	1.943026	0.054973	-0.363980436	33.86183526	-0.363980436	33.86183526
Groceries	85.41773139	29.68193925	2.877768	0.004947	26.49163171	144.3438311	26.49163171	144.3438311
Subscriptions	29.46294146	12.16960002	2.421028	0.017376	5.303230813	53.62265211	5.303230813	53.62265211

Appendix I:

Hypothesis Testing for Differences in Frequency of Groceries on Residence Type

Hypothesis testing: two-sided
H0:μRent-μDorm=0 Ha:μRent-μDorm!=0
Mean (Dorm)
3.93877551
Mean (Rent)
4.5
Mean Difference
0.561224489795918
Std^2 (Dorm)
8.100340136
Std^2 (Rent)
14.29591837
t-Statistics
0.8354817
Probability
>0.5(0.25*2)
Conclusion
Fails to reject h0 in favor of ha, meaning there is not enough evidence to conclude that there is a difference in the number of times students go to groceries based on different residence type