Milestone 1

Group Members: Kaitlyn Woolbert, Lavya Midha, Salma Salem, Jie Hu, Jasmine Zhou

1. Project idea: What is the main goal of your project? (KW)

We want to estimate the variables on which the average number of ride-share rides per person in a given municipality are affected by a number of factors relative to that municipality.

2. Variables: (KW, LM)

Dependent (response) variable is the average number of ride-share rides per person originated from each MA municipality in 2019.

Independent (predictor) variables are the variables you think may be useful in predicting the response. List at least 5 quantitative and at least 5 qualitative predictor variables. Your final model will most likely not include most of the variables you select. Examples of predictor variables:

a) Name of municipality - not a variable but a necessary identifier
b) Average income of municipality - quantitative
c) Date/month of ride – qualitative
d) Population of municipality - quantitative
e) Length of the ride (in miles) - quantitative
f) MA region from which ride originated - qualitative
g) total number of rides per municipality - quantitative
h) Speed of ride (in minutes) - quantitative
i) vehicle ownership percentage - quantitative
j) whether there is overnight parking availability in a municipality - qualitative data
k) company responsible for the ride-share - qualitative
l) company with highest ride-share percentage in municipality - qualitative data

3. Hypotheses:(LM, SS)

We hypothesize that the average ride-share rides per person originated from each municipality will increase as the population of the municipality increases. More densely packed cities will average higher ride-share overall. It is not a linear relationship between population and rides. More populated cities have more rides per city and less populated cities, probably partially due to the fact of less transportation access in general.

4. Contributions:

You must list how each member of the group contributed.

Lavya Midha: Came up with the topic of the project, brainstormed variables and helped write the hypothesis.

Kaitlyn Woolbert: brainstormed qualitative and quantitative variables for our study, and wrote the main goal of our project

Jie Hu: research data

Jasmine Zhou: research data

Salma Salem: hypothesis

Initials next to each question indicate which group member worked on that question

MILESTONE 2

Data collection method:

We are using a mix of data from a published source and data we will collect for our qualitative variables (information on availability of overnight parking, for instance). Most of our data can be found at https://tnc.sites.digital.mass.gov/. We will also be looking for a second source that provides us with qualitative data for the municipalities.

Attach sample data (only 5-10 observations).

We have not altered the data we found on the published source, but we intend to add more data to our spreadsheet once we narrow down our second source for qualitative data.

	TOWN	SUM_SQUARE_MILES	POP2010	ORIGIN_TRIPS_PER_PERSON	DESTINATION_TRIPS_PER_PE...	AVG_MILES_FROM_ORIGIN	AVG_MINS_FROM_ORIGIN	Type[1]	Type
314	WATERTOWN	4.13	31915	19.82	20.29	3.85	19.4	City	1
315	WAYLAND	15.85	12994	2.76	2.96	10.91	25.1	Town	0
316	WEBSTER	14.6	16767	0.64	0.69	12.34	21.34	Town	0
317	WELLESLEY	10.55	27982	10.63	11.36	6.07	17.64	Town	0
318	WELLFLEET	21.01	2750	2.8	3	5.11	11.28	Town	0
319	WENDELL	32.22	848	0	0.03	4.4	9.33	Town	0
320	WENHAM	8.14	4875	1.88	2.15	6.97	14.86	Town	0
321	WEST BOYLS...	13.86	7669	1.95	2.23	5.03	10.8	Town	0
322	WEST BRIDGE...	15.67	6916	3.2	3.47	5	10.67	Town	0
323	WEST BROOK...	21.11	3701	0.04	0.08	14.55	27.08	Town	0
324	WEST NEWBU...	14.73	4235	0.57	0.61	8.23	15.79	Town	0
325	WEST SPRING...	17.53	28391	3.92	3.96	3.29	8.63	City	1
326	WEST STOCK...	18.68	1306	0.09	0.09	7.84	15.04	Town	0
327	WEST TISBURY	26.29	2740	4.86	5.5	3.85	9.95	Town	0
328	WESTBOROUGH	21.45	18272	4.99	5.04	6.63	13.37	Town	0
329	WESTFIELD	47.31	41094	1.17	1.19	4.31	10.5	City	1
330	WESTFORD	31.36	21951	1.47	1.58	8.36	16.22	Town	0
331	WESTHAMPTON	27.36	1607	0.11	0.19	6	12.01	Town	0
332	WESTMINSTER	37.25	7277	0.45	0.57	9.52	17.05	Town	0
333	WESTON	17.33	11261	6.64	7.12	6.39	15.07	Town	0
334	WESTPORT	52.11	15532	0.43	0.36	7.21	12.9	Town	0
335	WESTWOOD	11.15	14618	7.24	7.8	6.14	13.62	Town	0
336	WEYMOUTH	17.83	53743	5.71	5.74	3.95	10.55	City	1
337	WHATELY	20.67	1496	0.5	0.63	8.22	14.67	Town	0
338	WHITMAN	6.96	14489	2.35	2.39	4.23	10.71	Town	0
339	WILBRAHAM	22.33	14219	1.02	1.12	4.71	11.65	Town	0
340	WILLIAMSBURG	25.67	2482	0.32	0.51	5.72	13.37	Town	0
341	WILLIAMSTOWN	46.85	7754	0.07	0.1	5.28	11.29	Town	0
342	WILMINGTON	17.14	22325	3.55	3.74	5.79	12.72	Town	0
343	WINCHENDON	44.11	10300	0.02	0.07	13.81	23.07	Town	0
344	WINCHESTER	6.35	21374	5.72	6.07	4.79	14.18	Town	0

4.

Contributions:

Kaitlyn Woolbert: helped make the corrections to milestone 1, continuing to search for a second qualitative source

Lavya Midha: brainstormed ideas for qualitative variables that can be included in the future, help edit the milestone.

Jie Hu: Help search for a qualitative source, brainstorming with the qualitative variables

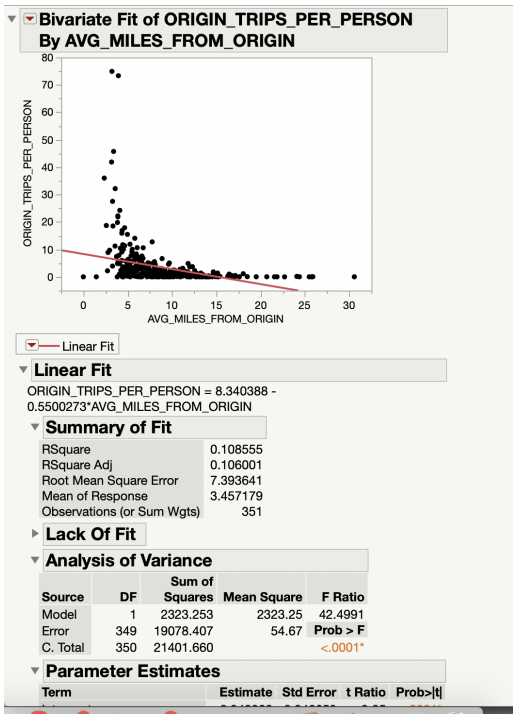
Jasmine Zhou: Research data based on the topic, organizing qualitative data.

Salma: researched availability of qualitative data

MILESTONE 3

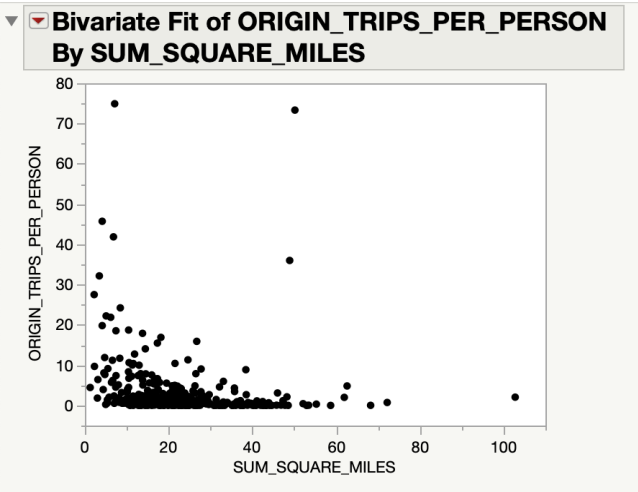
Scatter plots: For all your quantitative covariates, X_1, X_2, …, X_k, plot the individual Fit Model with Y as the response and X_i as the predictor. Identify outliers and decide if you need to do anything about them.

Salma: average miles from origin

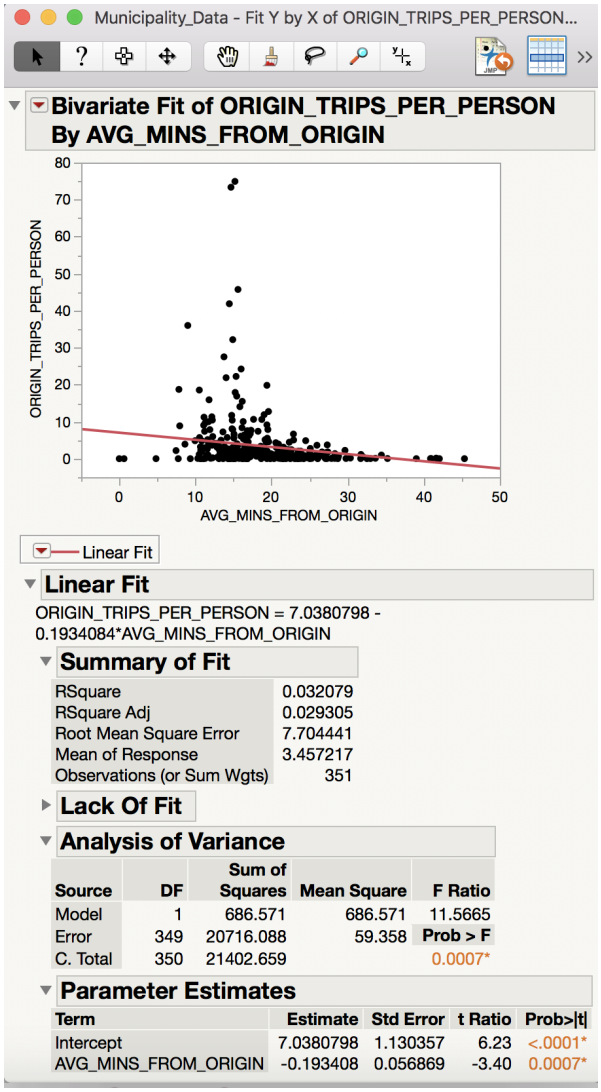


There doesn't seem to be a relationship between average miles from origin to origin trips person as the R square value is low. There are 3 outliers and it may help the data if they are removed.

Kaitlyn: sum square miles

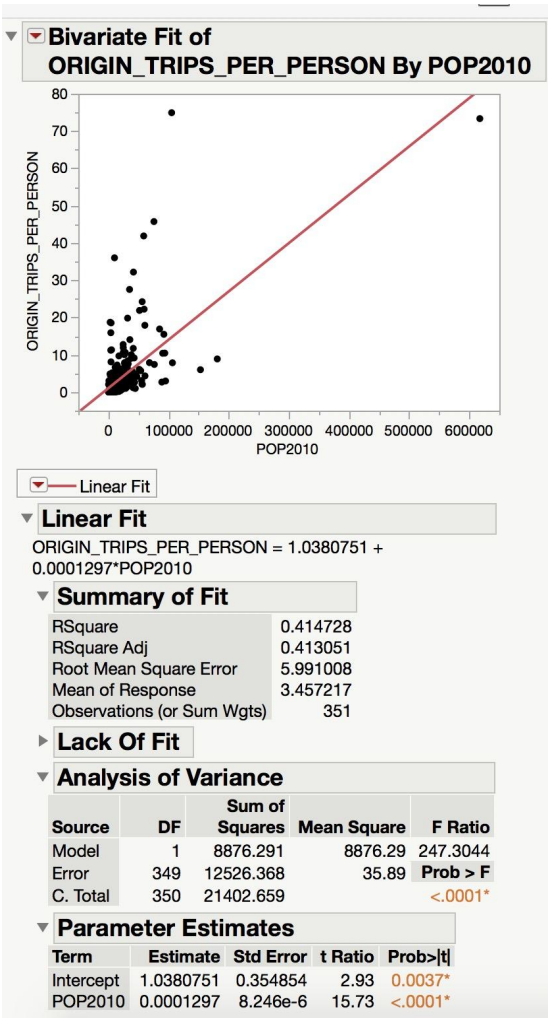


Jie: average mins from origin



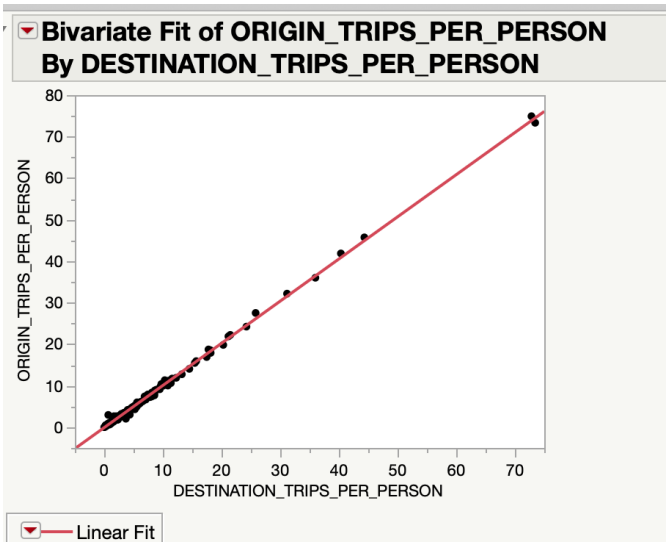
It seems like there is no relationship between the avg_mins from origin and origin trip per person, the R squared is small. I would try to do some transforming to fix it.

Jasmine: population of 2010



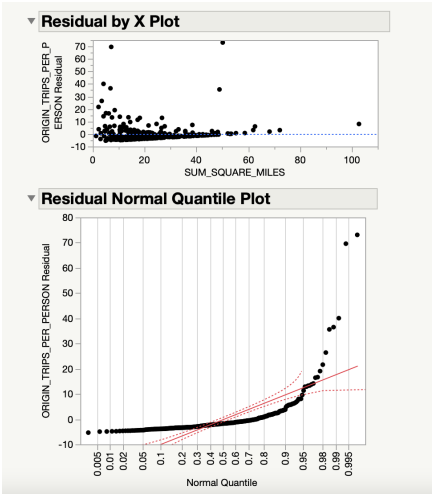
There is no significant relationship between the population of 2010 and origin trips per person; however, as we can see that datas are gathered together, we can conclude that most individual have their origin trips in the range between 0 to 15.

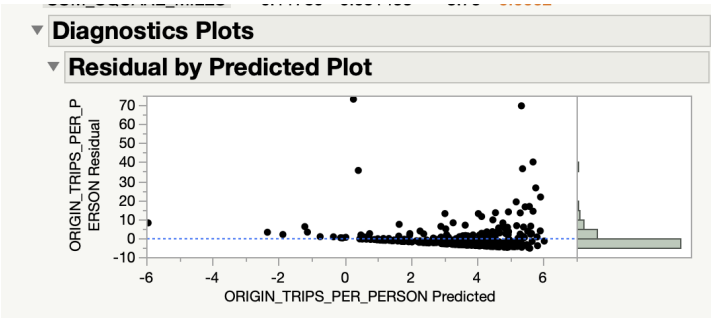
Lavya: destination trips per person



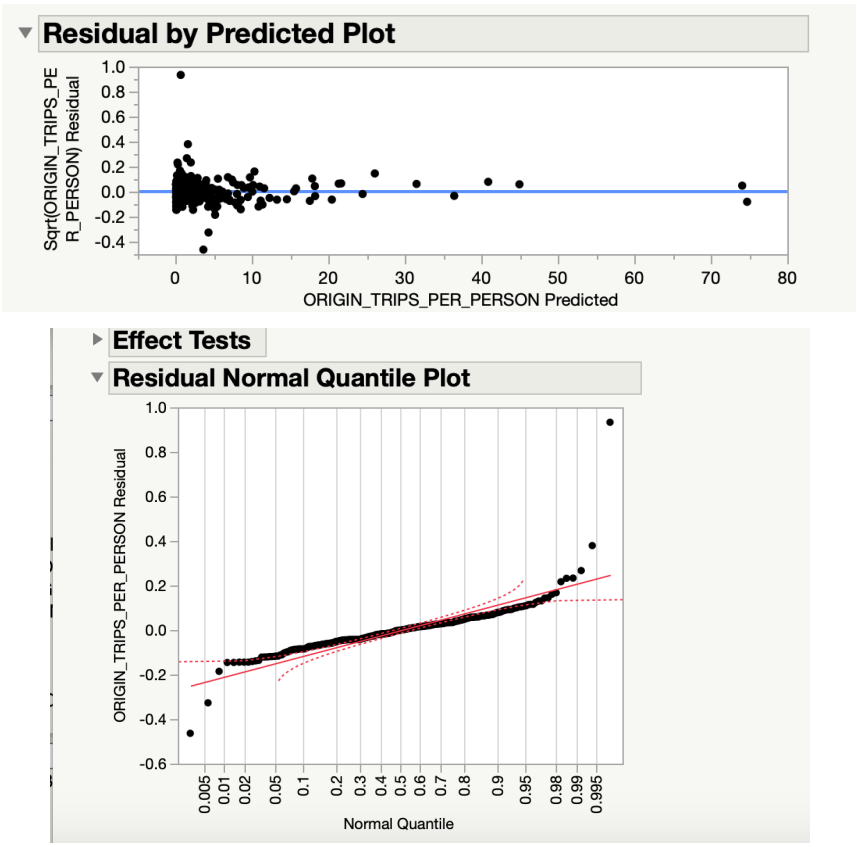
Analyze residuals: For each single linear regression model of Y vs X_i, plot the residuals. Are the assumptions of a linear regression model satisfied (residuals random and lie on the line in the normal quantile plot)? Do you need to transform any of the variables? Should you consider removing any variables?

Kaitlyn: The assumptions are not satisfied for this relationship; after trying all the available transform options, there is still not any strong relationship between the two variables. Removal of the outliers (row 239, 36, 50, 197) does not give us a closer relationship.

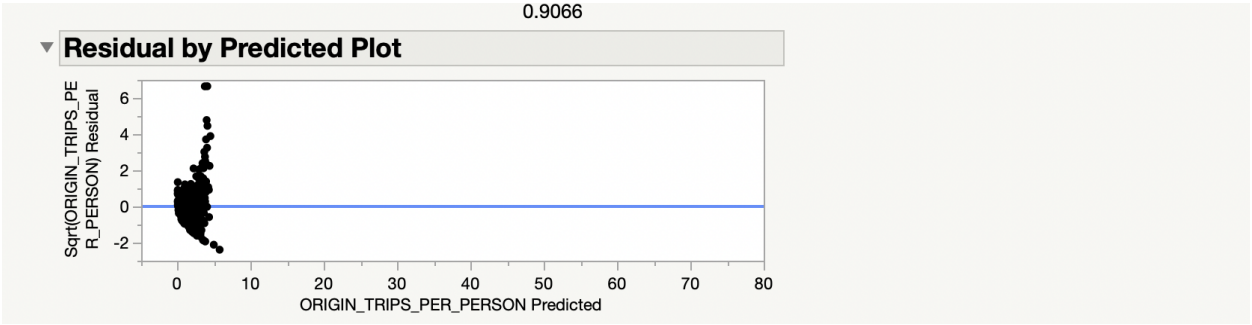




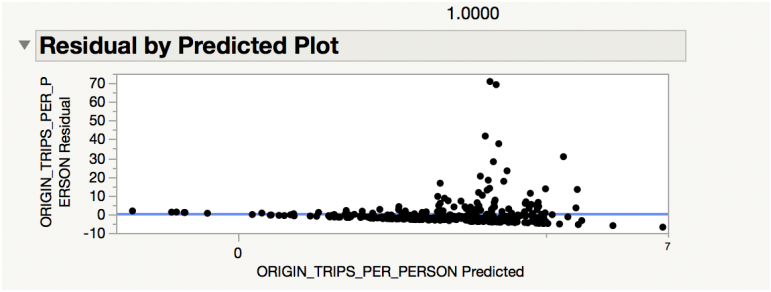
Lavya: If we transform the variables to square root we get a stronger relationship, if we see the normal quantile plot. However removal of outliers (36,50) does not improve the problem.

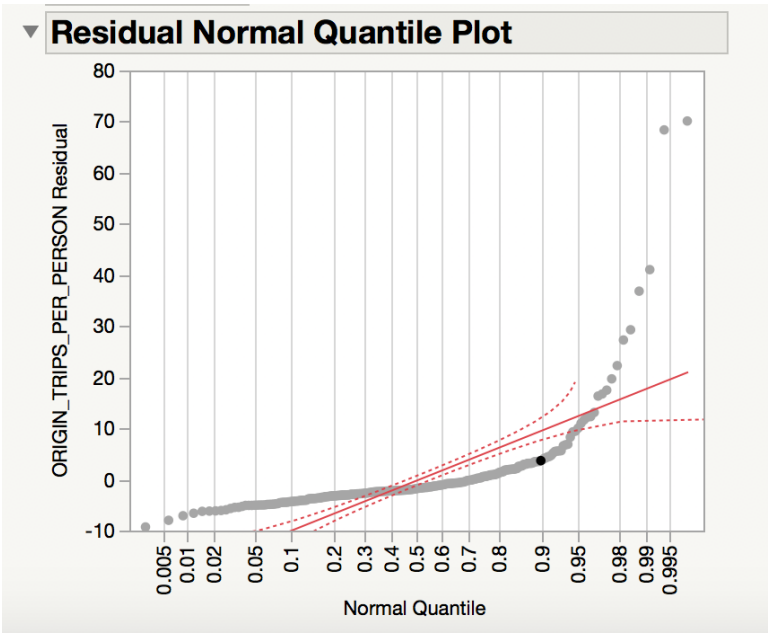


The assumptions of a linear regression model are not satisfied. The residuals are not randomly distributed.

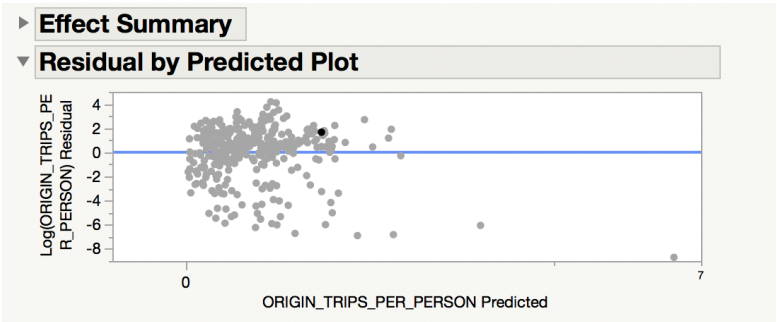


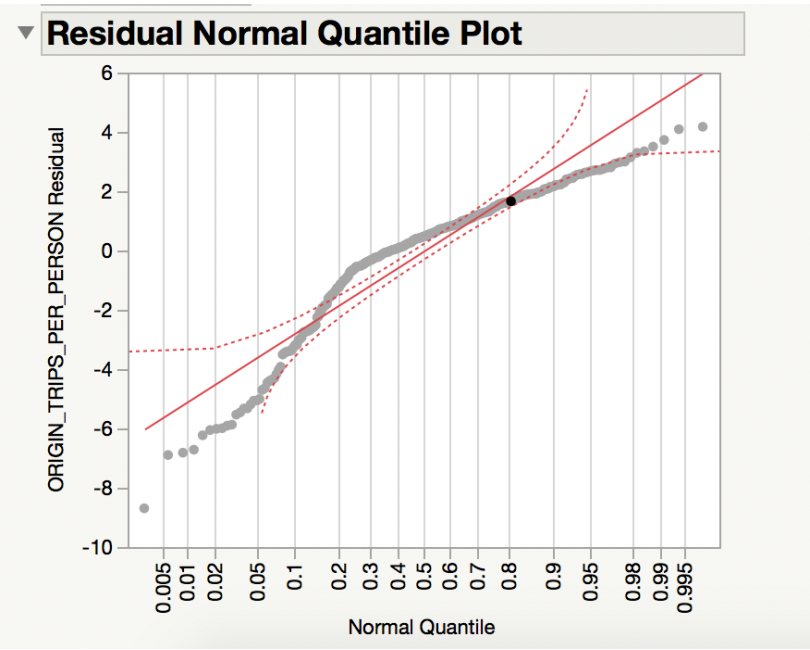
Jie: The assumption of a linear regression model is not satisfied, the points are not random and lie on the line in the normal quantile plot. After transforming the variable to the Log, we get a little bit stronger relationship (but still not very strong relationship). Removing the outliers (rows 129 156 253) does not solve the problem.



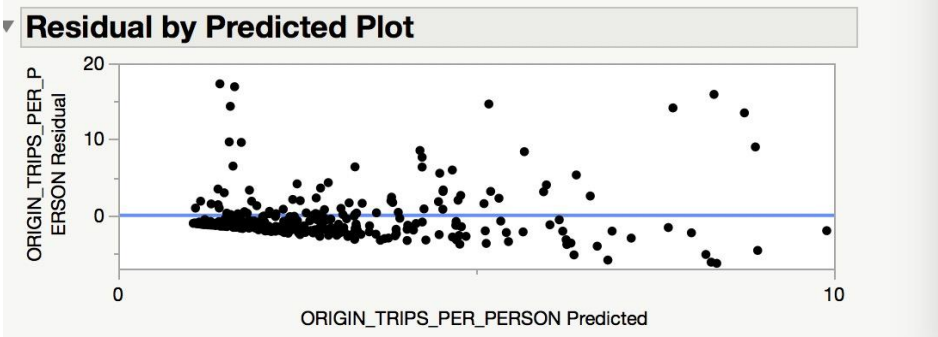


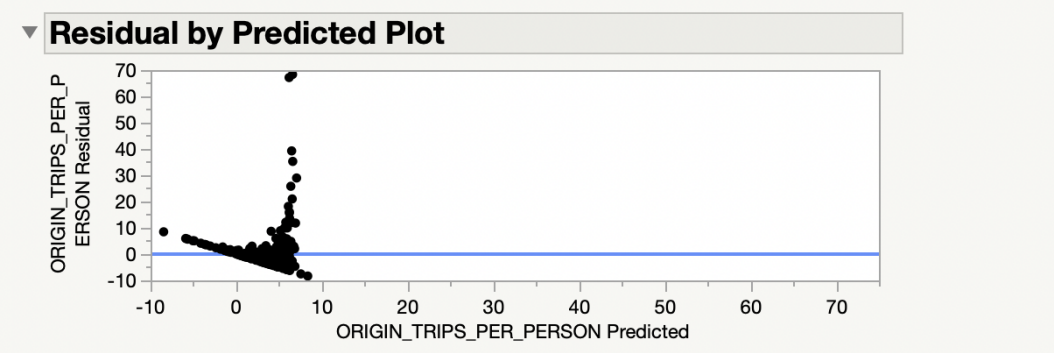
Transform to the Log:

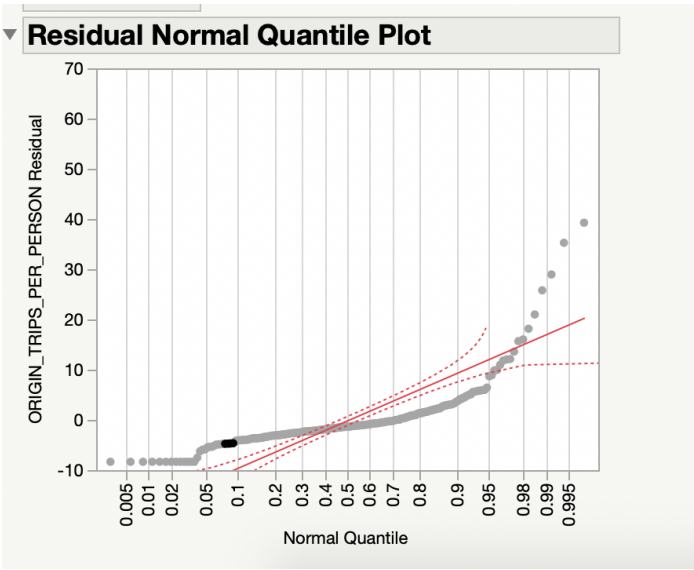




Jasmine Zhou : The assumptions are not satisfied for this relationship between origin trips per person and population of 2010. Though a transformation might be useful, however there are too many outliers in the plot. Removing some of them does not solve the problem and it did not show a stronger reltionship.







Salma: The assumption of a linear regression model is not satisfied as the points are not distributed equally randomly and removing the outliers does not seem to have a ibeneficial effect on the data. Also the points don't seem to fall along the red line on the normal quantile plot further showing this model does not satisfy the assumption of a linear regression model. However transforming the data, does seem to hint at a negative relationship between average miles from origin and origin trips per person.

At the end we can conclude we would be able to conclude that deleting outliers where the origin trips per person are zero (since it would not matter) and where average miles per person are zero (in order to perform reciprocal transformation) would be best for our data.

4. Optional: If you have any problems, list them.
5. **Contributions: Describe how each member contributed to the project.**

Kaitlyn: analyzed relationship between sum square miles and origin trips per person
Jie: analyzed relationship between average mins from origin and origin trips per person
Jasmine: analyzed relationship between population of 2010 and origin trips per person
Lavya: analyzed relationship between destination trips per person and origin trips per person

Salma: analyzed relationship between average miles from origin and origin trips per person

Multicollinearity: Check the quantitative covariates for multicollinearity and comment on any correlations.

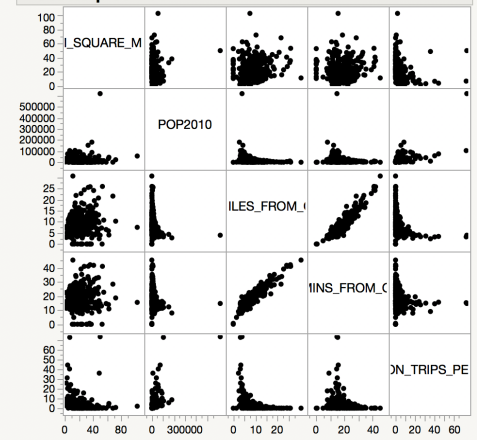
Multivariate

Correlations

	SUM_SQUARE_MILES	POP2010	AVG_MILES_FROM_ORIGIN	AVG_MINS_FROM_ORIGIN	DESTINATION_TRIPS_PER_PERSON
SUM_SQUARE_MILES	1.0000	0.0763	0.1882	0.0475	-0.2007
POP2010	0.0763	1.0000	-0.2403	-0.1495	0.6475
AVG_MILES_FROM_ORIGIN	0.1882	-0.2403	1.0000	0.9379	-0.3283
AVG_MINS_FROM_ORIGIN	0.0475	-0.1495	0.9379	1.0000	-0.1746
DESTINATION_TRIPS_PER_PERSON	-0.2007	0.6475	-0.3283	-0.1746	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix



There is in between mins from origin and miles from origin.
So, we decide to remove average mins from origin.

Model fit: Check the full model with all the covariates, including qualitative variables, to determine if it is statistically significant, using any transformations that you determined to use in Milestone 3. Give the ANOVA table, state appropriate hypotheses, conclusions, R^2, and any other important information about the full model.

Summary of Fit

RSquare	0.998445
RSquare Adj	0.998421
Root Mean Square Error	0.321607
Mean of Response	3.76854
Observations (or Sum Wgts)	322

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	20991.149	4198.23	40589.70
Error	316	32.684	0.10	Prob > F
C. Total	321	21023.833		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.221287	0.057708	-3.83	0.0002*
AVG_MILES_FROM_ORIGIN	0.0016856	0.004746	0.36	0.7227
POP2010	-2.405e-6	6.474e-7	-3.72	0.0002*
SUM_SQUARE_MILES	0.003756	0.001521	2.47	0.0141*
Type	0.2569577	0.054886	4.68	<.0001*
DESTINATION_TRIPS_PER_PERSON	1.0206271	0.003185	320.43	<.0001*

Summary of Fit

RSquare	0.995553
RSquare Adj	0.995483
Root Mean Square Error	0.086457
Mean of Response	1.455631
Observations (or Sum Wgts)	322

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	528.83479	105.767	14149.67
Error	316	2.36206	0.007475	Prob > F
C. Total	321	531.19685		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.073415	0.015604	-4.70	<.0001*
Reciprocal(AVG_MILES_FROM_ORIGIN)	0.3087897	0.088623	3.48	0.0006*
Sqrt(DESTINATION_TRIPS_PER_PERSON)	0.9915572	0.00599	165.53	<.0001*
Type	0.0537517	0.014994	3.58	0.0004*
POP2010	1.3985e-7	1.624e-7	0.86	0.3898
SUM_SQUARE_MILES	-1.162e-5	0.000421	-0.03	0.9780

$H_0 : \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6$

H_a : at least one of the β_i is non zero.

Given that Prob>F= <0.0001 which is less than 0.05, therefore there is enough evidence 9that the model is useful and at least has one variable which is linearly related to our dependent variable.

$R^2 = 0.9958$

Which gives us that 99.58% of the total sample variability around y_bar that is explained by the linear relationship between the independent variables and the dependent variables.

We would use the model without transformations since the R^2 it is almost the same.

4. Variable Selection: Try several variable selection techniques. Only report on the final model you decide to use and how you selected the variables. Be sure the analyze the residuals and check for normality.

We did a stepwise regression using the Minimum AIC and forward selecting method.

After analyzing residuals and checking for normality, we felt justified in including the four variables our stepwise regression model selected; Type, population, sum squared miles and destination trips per person. As these variables clearly had the lowest p values and residuals that seemed less patterned, it seemed appropriate to include only these in our final model, which is posted below.

29 rows not used due to excluded rows or missing values.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
32.697214	317	0.3211632	0.9984	0.9984	4.1261554	5	189.5648	211.9455

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	-0.2063955	1	0	0.000	1
☐	☑	SUM_SQUARE_MILES	0.00386621	1	0.697105	6.758	0.00977
☐	☐	AVG_MILES_FROM_ORIGIN	0	1	0.013048	0.126	0.72269
☐	☑	DESTINATION_TRIPS_PER_PERSON	1.02041129	1	11016.11	106801.3	0
☐	☑	POP2010	-2.4137e-6	1	1.439629	13.957	0.00022
☐	☑	Type	0.25181216	1	2.340113	22.687	2.91e-6

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	DESTINATION_TRIPS_PER_PERSON	Entered	0.0000	20987.98	0.9983	28.615	2	213.021	224.27	○
2	Type	Entered	0.0003	1.477005	0.9984	16.335	3	201.525	216.497	○
3	POP2010	Entered	0.0024	0.979406	0.9984	8.866	4	194.281	212.964	○
4	SUM_SQUARE_MILES	Entered	0.0098	0.697105	0.9984	4.1262	5	189.565	211.945	○
5	AVG_MILES_FROM_ORIGIN	Entered	0.7227	0.013048	0.9984	6	6	191.526	217.591	○
6	Best	Specific	.	.	0.9984	4.1262	5	189.565	211.945	●

5. Interpret the Model: Do the coefficients have a natural interpretation? Plug in some values for the predictor variables to determine the expected response value and individual prediction intervals. Are the results what would be expected?

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.206396	0.039599	-5.21	<.0001*
SUM_SQUARE_MILES	0.0038662	0.001487	2.60	0.0098*
DESTINATION_TRIPS_PER_PERSON	1.0204113	0.003122	326.80	<.0001*
Type	0.2518122	0.052867	4.76	<.0001*
POP2010	-2.414e-6	6.461e-7	-3.74	0.0002*

(avg number of origin trips per person: 3.77

The coefficients show a strong relationship between type of municipality, destination trips per person, sum square miles, and population of 2010 on origin trips per person as it is very close to 1 being 0.998. For type of municipality cities have a greater number of origin trips per person as compared to towns. For typical inputs the response would be about 3.77 as that is the average origin trips per person based on the data set. Increasing sum square miles of the municipality and destination trips per person causes the response variable to increase while increasing the population variable causes the response variable to decrease. If we include the municipalities where the origin trips were 0 then the response variable should be 0 but it isn't. The model is only for municipalities where the origin trips weren't 0 in the data set because those values were excluded. Applying to other municipalities would be extrapolation as the final model only gives insight on municipalities in Massachusetts.

Overall after plugging values in the model is a positive function with increasing destinations trips per person.

How did transformations affect the model? When we transformed the data the R^2 was 0.995

If we don't don't transform the data the R^2 is 0.998

6. Contributions: Describe how each member contributed to the project.

Lavya Midha : Helped complete stepwise regression and make the anova table

Kaitlyn Woolbert: helped complete stepwise regression/variable selection, wrote about data collection, Why We Picked Variables, Background, Issues and Hypotheses

Jasmine Zhou: Helped complete transformation and stepwise regression

Salma Salem: interpreted the model

Jie Hu: check the multicollinearity