TERRA R. EDENHART-PEPE
WEEK 7
10.26.18

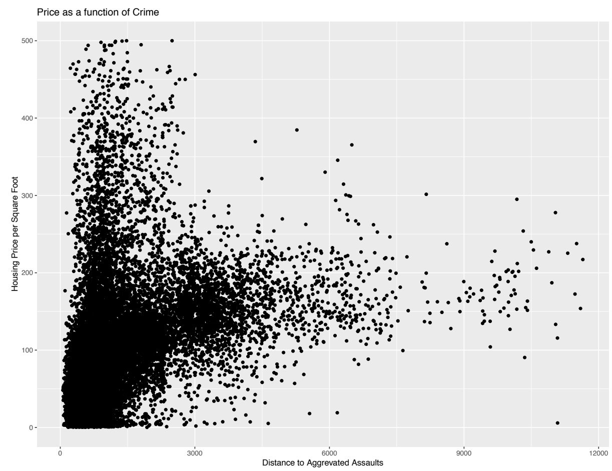
ANALYSIS 1 Price per squarefoot

and explanatory variables

the kitchen sink

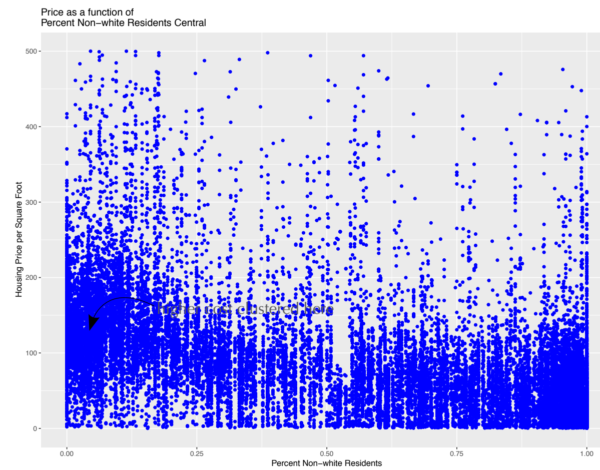
Crime
(in distance units to aggrevated assaults)

The scatterplot shows a price pattern in which the highest prices exist in the areas of lowest crime, specifically lowest rate of reported aggrevated assault. None of the highest crime areas yield homes with the highest price per square foot.



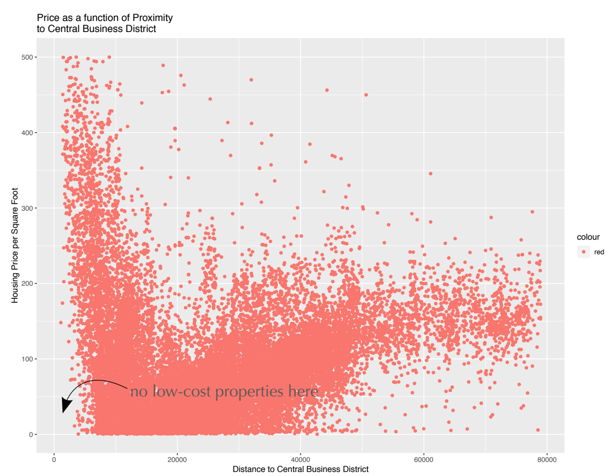
Percent Non-white

I initially included this variable under the (false) assumption that percent non-white would be a proxity for multiple other factors (highly, negatively correlated with income) delivering a “two-in-one” variable. Retrospectively, (after completing the homework), I understand that this two-in-one variable is exactly the type that should be avoided, as such variables “muddy the water,” e.g. especially because it overlaps with the crime variable creating a weighted model and bias.
While there is a lot of noise in this plot, it illustrates a non-linear pattern in which populations heavily non-white or heavily white are clustered. The lowest percent non-white yields the higher price per square foot.



Proximity to Central Business District
(in distance units are feet)

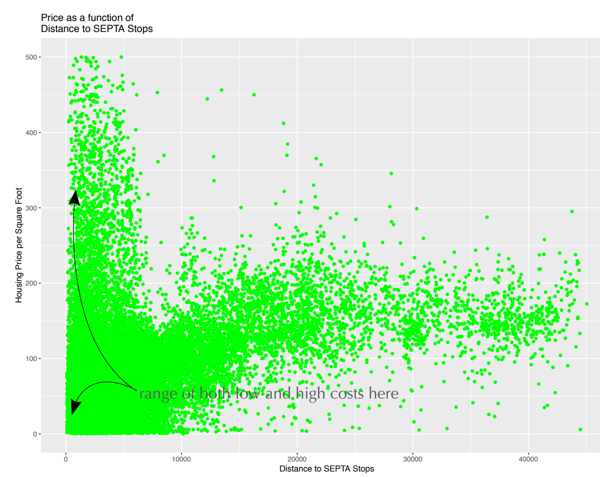
The intervals at which proximity seems to have a strong impact are the close range (<20,000) and mid-range (<45,000). It is plausible that price per square feet is generally at it’s highest in areas within 10,000 because the development density limits supply causing a premium increase on space. This analysis makes more sense when the units change. Cost increase associated with feet is very small but associated with 1000 is a $amount that can be conceptualized relative to total houseing price.



Distance to SEPTA Stops
(in distance units are feet)

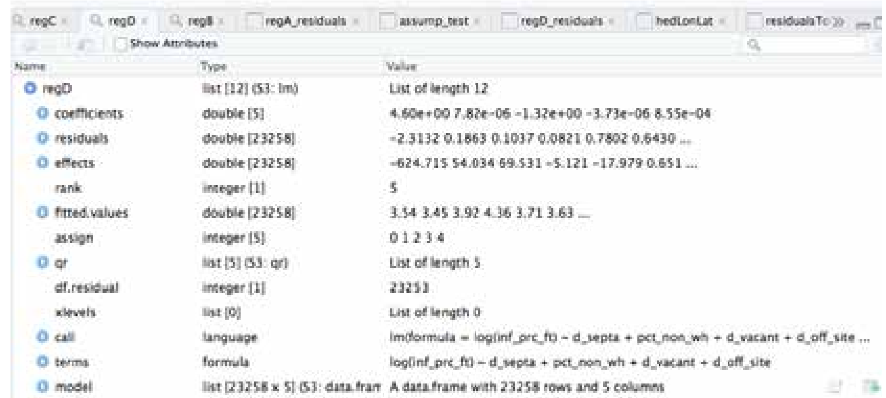
The greatest density of cases exists in the low to mid-range cost areas, because that is the majority of the available properties. Again, higher costs exist closer to the target (transit stops).

Transit stops were strategically sited in areas that had commerce, which means that the areas around transit stops are already high value. These are complex areas containing many factors; there could be other factors which have a stronger relationship to price.



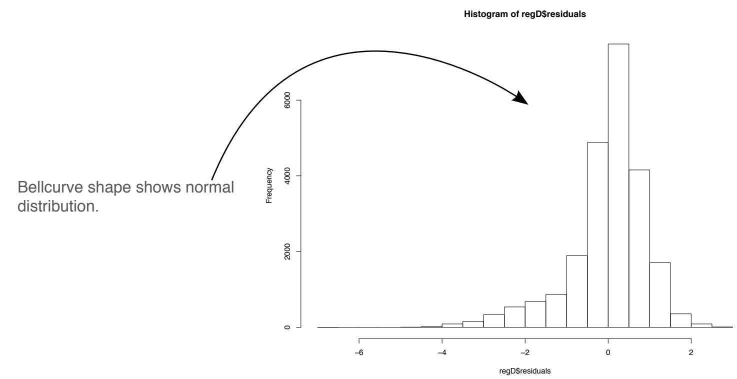
ANALYSIS 1

the kitchen sink



Name	Type	Value
regD	list [12] (S3: lm)	List of length 12
coefficients	double [5]	4.60e+00 7.82e-06 -1.32e+00 -3.73e-06 8.55e-04
residuals	double [23258]	-2.3132 0.1863 0.1037 0.0821 0.7802 0.6430 ...
effects	double [23258]	-624.715 54.034 69.531 -5.121 -17.979 0.651 ...
rank	integer [1]	5
fitted.values	double [23258]	3.54 3.45 3.92 4.36 3.71 3.63 ...
assign	integer [5]	0 1 2 3 4
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	23253
xlevels	list [0]	List of length 0
call	language	lm(formula = log(inf_prc_ft) ~ d_septa + pct_non_wh + d_vacant + d_off_site ...
terms	formula	log(inf_prc_ft) ~ d_septa + pct_non_wh + d_vacant + d_off_site
model	list [23258 x 5] (S3: data.fran	A data.frame with 23258 rows and 5 columns

Screenshot: Final Kitchen Sink Regression

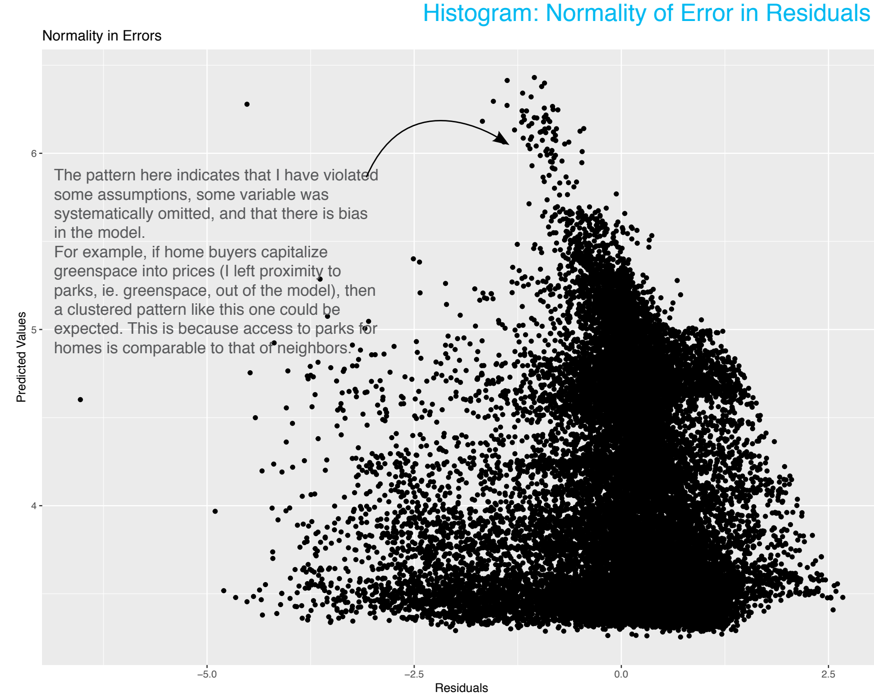


Histogram: Normality of Error in Residuals

Distance to SEPTA coefficient indicates practical significance or actual value of the relationship between the variables. Given 1 unit (foot) of change in proximity, there will be an expected change in price of .00001651.

28% of variation in response variable can be explained by variation in explanatory variable. Ratio of how much variation is taken up by model/ total variation. .2882 is reasonably high but domain knowledge determines whether this value meets expectations of “good”.

The residual standard error represents the typical error in this model and quantifies how well or poorly the model is performing (or how well/ poorly it is predicting data on average). This model is off by .9 when it tries to predict data, which is “good”.



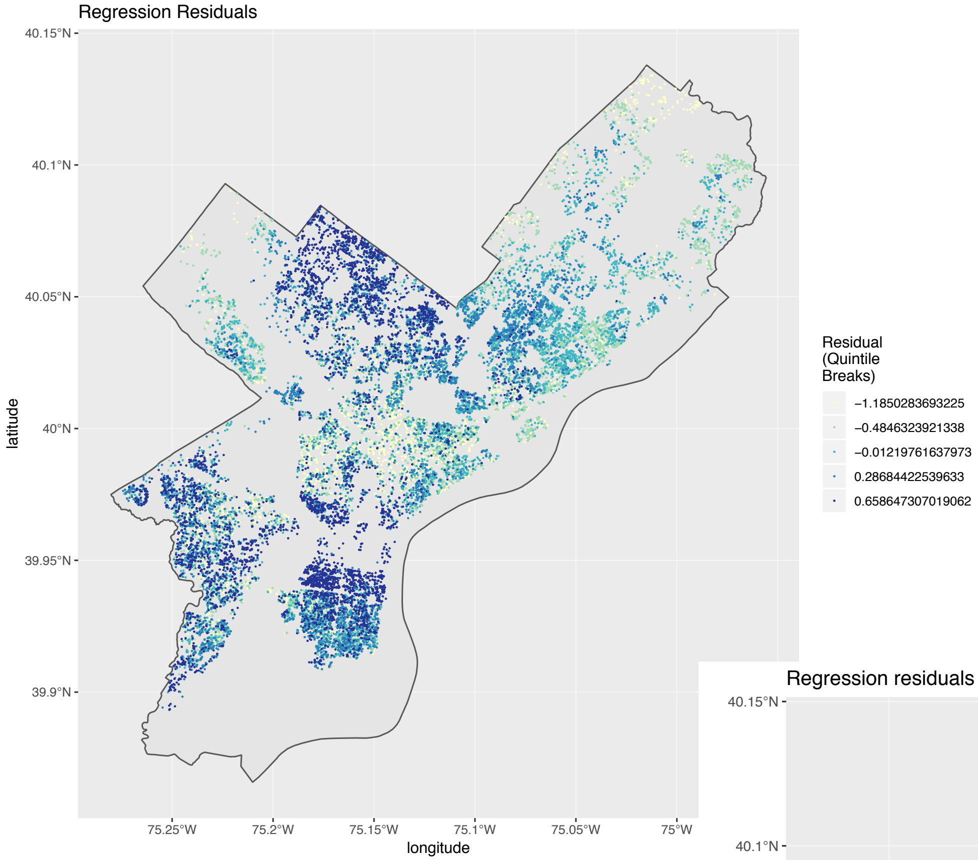
Homoskedasticity: Amount of Variation

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.5287 -0.3160  0.1413  0.5470  2.6730

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.670e+00  1.987e-02  235.07   <2e-16 ***
d_septa      1.651e-05  1.608e-06   10.27   <2e-16 ***
pct_non_wh  -1.226e+00  1.901e-02  -64.51   <2e-16 ***
d_crime      1.468e-04  8.551e-06   17.17   <2e-16 ***
d_cbd       -7.701e-06  6.575e-07  -11.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9296 on 23253 degrees of freedom
Multiple R-squared:  0.2883,	Adjusted R-squared:  0.2882
F-statistic:  2355 on 4 and 23253 DF,  p-value: < 2.2e-16
```

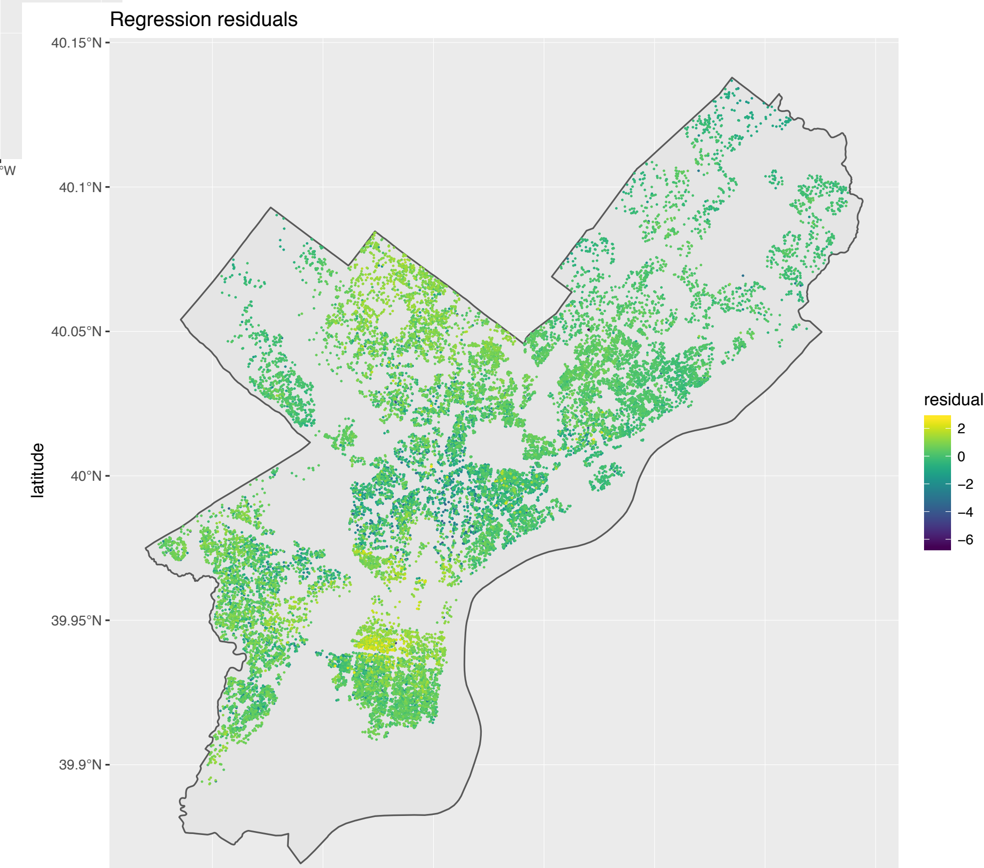
ANALYSIS

1

the kitchen sink




```
Moran I test under randomisation

data:  regA$residuals
weights: nb2listw(spatialWeights, style =
"W")

Moran I statistic standard deviate = 112.27,
p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation
Variance
     4.870658e-01     -4.299781e-05
1.882500e-05
```



Not all the diagnostic tests yield results that confirm the performance of the model and significance of the variables. The clear clustering of residuals and clustering in the Homoskedasticity test are the biggest red flags, indicating bias and a systematic omission of variables. It may be likely that a variable that is omitted is the tendency for people to self-select locations of residence, a variable that can not be quantified.

ANALYSIS

2

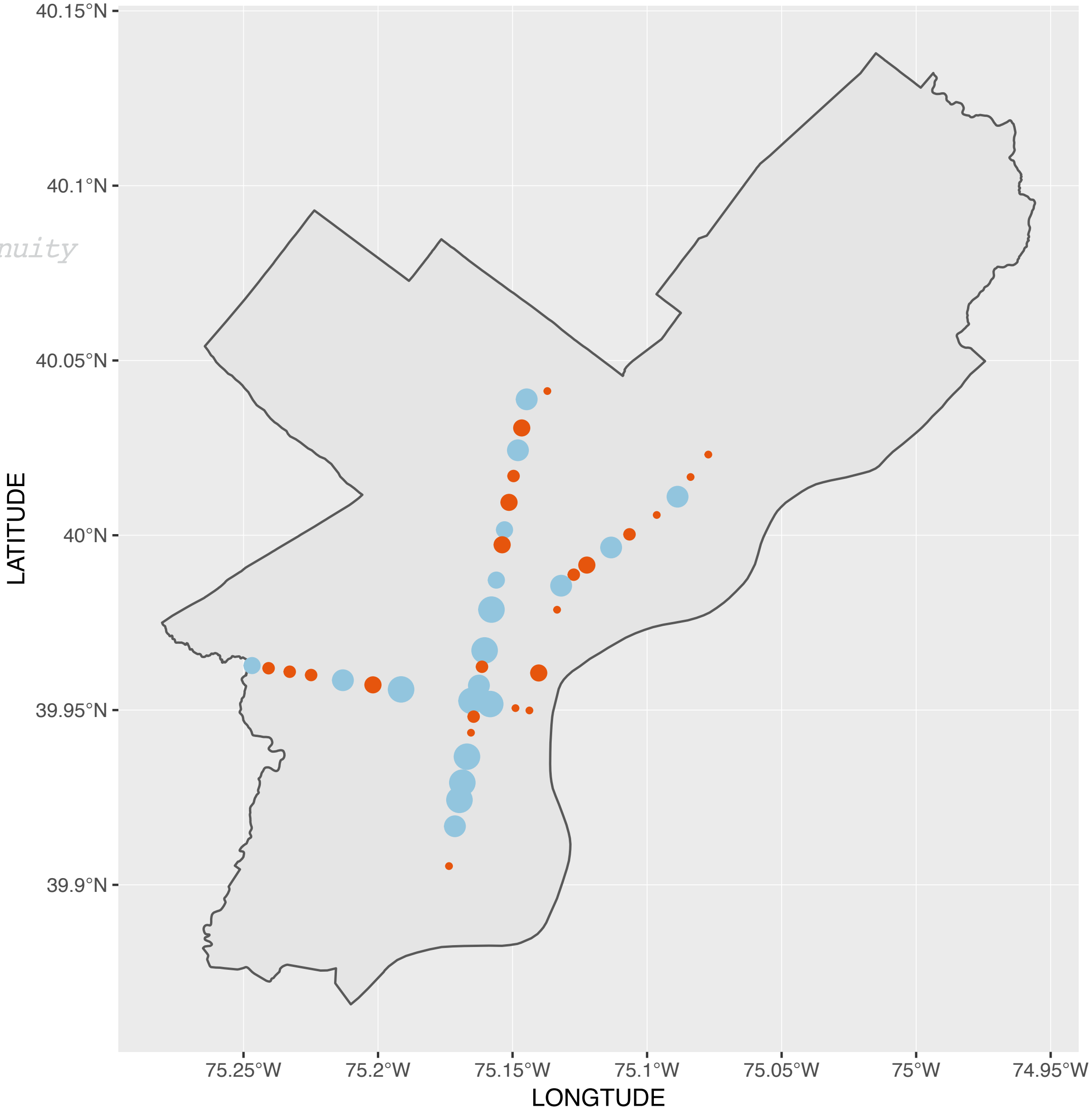
the transit discontinuity

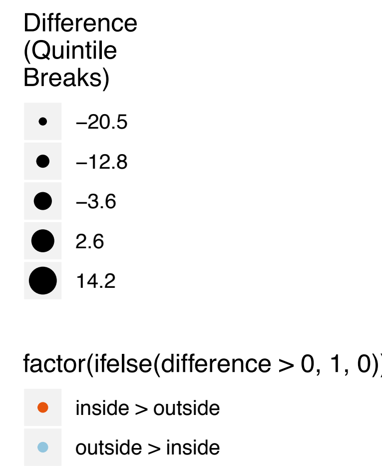
Contemporary transit planners suggest that residents are willing to walk no more than one quarter of a mile to a subway station. Instead, they will drive or find an alternative means of transport. This idea provides an interesting "quasi-experimental" research design opportunity where we can test for differences in home prices on either side of the quarter-mile boundary.

While this theory may be appropriate as a test, the acceptable distance assumption may not necessarily true in all areas. Anecotally speaking, it seems that culture may changing to be more mobile (bike-centered), urban culture, in which this rule does not necessarily apply. To improve the rigor of this quasi-experimental research, beginning with primary social data collection regarding current mobility trends could improve accuracy. However, generally speaking this research design makes sense and adjustments may simply be in the form of a wider diameter circle.

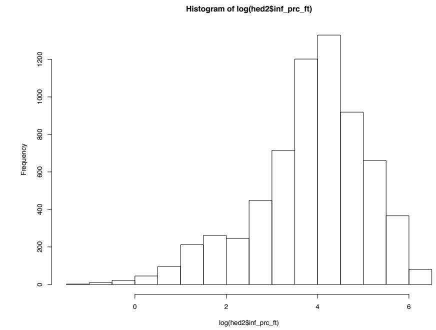
Does the specific station matter?

Or is price a function of general proximity to transit?





There appears to be a relatively equal distribution of preference, suggesting that individual stations are not much more desireable than others. I would interpret that pattern to mean that proximity is the key decision factor and buyers are more willing to pay for transit, in general.



Histogram: LOG of price/sf

ANALYSIS 2

the transit discontinuity

Station	1. Just the fixed effect	2. With station fixed effects	3. Distance to Parks
	Dependent variable: log(inf_prc_ft)		
40TH STREET		-1.5877***	
		-0.3829	
46TH STREET		-1.3493***	
		-0.3847	
52ND STREET		-2.2064***	
		-0.3737	
56TH STREET		-2.2627***	
		-0.377	
60TH STREET		-2.3030***	
		-0.3753	
63RD STREET		-1.9060***	
		-0.3795	
ALLEGHENY		-2.1575***	
		-0.3704	
BERKS		-1.3154***	
		-0.3736	
CECIL B MOORE		-2.0553***	
		-0.3799	
CHURCH		-1.8937***	
		-0.3815	
ELLSWORTH-FEDERAL		-1.0866***	
		-0.3723	
ERIE		-2.4807***	
		-0.3749	
ERIE-TORRESDALE		-1.4154***	
		-0.3757	
FAIRMOUNT		-1.1919***	
		-0.3843	
FERN ROCK T.C.		-1.4779***	
		-0.3768	
FRANKFROD T.C.		-1.3789***	
		-0.3732	
STATIONHUNTING PARK		-2.5144***	
		-0.3832	
STATIONHUNTINGDON		-2.5218***	
		-0.375	
STATIONLOGAN		-1.9561***	
		-0.3768	
STATIONMARGARET-ORTHODOX		-1.8744***	
		-0.3765	
NORTH PHILADELPHIA		-3.0459***	
		-0.3796	
OLNEY		-1.5325***	
		-0.3861	
SOMERSET		-2.5338***	
		-0.3724	
SUSQUEHANNA-DAUPHIN		-2.7773***	
		-0.3727	
TASKER-MORRIS		-1.2568***	
		-0.3716	
TIOGA		-1.8578***	
		-0.373	
WYOMING		-2.3517***	
		-0.377	
YORK-DAUPHIN		-2.4767***	
		-0.3782	
lt_qrtMi	-0.0879***	0.0163	
	-0.0309	-0.0249	
d_parks			-0.0004***
			-0.00003
Constant	3.8867***	5.5343***	4.2642***
	-0.0196	-0.3685	-0.0335
Observations	6,612	6,612	6,612
R^2	0.0012	0.379	0.0279
Adjusted R^2	0.0011	0.375	0.0278
Residual Std. Error	1.2320 (df = 6610)	0.9745 (df = 6569)	1.2154 (df = 6610)
F Statistic	8.0760*** (df = 1; 6610)	95.4502*** (df = 42; 6569)	189.8091*** (df = 1; 6610)

Regression 1: Fixed Effect
This test asked if there is an average difference in price on either side of the buffer, assuming all else equal, the average log price difference between home sales within the quarter mile buffer and those just outside is -8.79%.

Regression 2: Fixed Effects with Stations
This test describes the phenomena of a given station and is conditional on the closest station of each sale, the log price difference between home sales within the quarter mile buffer, and those just outside is variable by station. For this regression, each station is a variable, which has a PValue. Approximately half of the stations were removed from the chart to the left for having standard errors.

Regression 3: Distance to parks
Distance to parks PValue is the lowest of the values, indicating the highest statistical significance. It may be assumed that proximity to parks exerts and influence on cost of realestate.

What did these additional station controls do to the transit fixed effect coefficient?
The additional station controls allowed the model to examine the inside/outside buffer effect by withholding potential preference within different areas of the city. Because the PValue of the percent per squarefoot with the individual stations controlled is a lower value, the confidence level that proximity to transit is higher.

The R^2 value increased:
1. .00106
2. .375
3. .027

A possible rease for increase is colinearity amongst controls. If the variables input in the model are correlated with the quarter mi. transit fixed effect or eachother, the model is biased.

How well is the model doing? When it tries to use proximity to transit to predict price the model is off by 1.2, .9, or 1.2.

What is the willingess to pay that you have estimated? One unit of change (in distance outside of the quarter mile zone) the price per square foot decreases by .0879 units, or every 1000 feet outside of the 1/4 zone, the price per square feet increases $87.9. The magnitude of this relationship is great compared to the average price per square foot, in a 1,500 sf home, this would equate to an addition $131,850.

Does this research design help identify the willingness to pay for transit? This research design helps to identify the willingness to pay by controlling for colinear factors that would impact the results.

The two things that are not included in this data or tests are self-selection phenomena and individual, interior real estate information (“specs”), both of which have a significant impact on the amount a buyer is willing to pay.