MWSUG BL-104
EXPLORE YOUR DATA BEFORE YOU RUSH TO ANALYSIS, YOU WILL THANK ME LATER: EXPLORATIONS IN CROSS SECTION DATA

Steven C. Myers
Department of Economics, College of Business Administration
The University of Akron, Akron, OH

ABSTRACT

Economists, business leaders and analysts spend a great deal of time analyzing structured cross sectional data. This paper is an introduction to exploratory data analysis for economic and business data analytic students in an introductory course in economics to teach data handling and SAS programming and features SAS® PROCS MEANS, UNIVARIATE, SGPLOT, FREQ, TABULATE, CORR, TTEST and REG. A data set on rents paid is used to illustrate the solution of the problem: do female students pay higher rents than male students?

It is essential to learn all you can about your data before rushing to analysis, yet analysts typically rush to more advanced and fancier techniques by ignoring or giving only cursory concern to the underlying data. In this paper we show how to ground the analysis in a firm understanding of the data generating process and suggest many ways to learn about the under lying data.

INTRODUCTION

Data takes time. Hasty analysis can lead to all sorts of problems. Hasty data work is defined as a combination of ill-considered problem articulation, insufficient data cleaning and other preparation and poorly developed modeling. Often this is combined with insufficient groundwork for statistical inference leading to unreliable conclusions. Hasty data work is tempting because of your curiosity about what the data says and the pressure of getting to an answer quickly. But data takes time.

It is this time component that I want to emphasize in this paper.

One data scientist recently remarked that when directing his team to consider a new problem, he denies them access to data until they have considered the problem theoretically and solved it through a thought process which identifies the features (or data variables) needed.[1] Once that first step is completed, the next step is acquiring the data variables that match their needs. This is not always possible and even the variables that may match their needs may not be measured as theoretically pure as one would like. Contrast the need for a variable measuring ability or motivation with the test scores you may have. How well the test score represents the needed ability measure may influence confidence in the analysis and interpretation of results needs to take account of the potential biases of the use of the data proxy. Some variables may not be possible to measure or proxy and will represent missing relevant observations leading to potential omitted variables bias, again which must be acknowledged when interpreting and using the results.

Analysis cannot begins until after the data has been acquired and should never begin until the data has been cleaned. Much data is dirty or messy and the process of cleaning it may take many forms. Once cleaned, the data is ready for analysis, but may need further transformation to fit our model.

So all investigation must fulfill three areas. (1) The problems and questions that arise from the business leaders (2) good foundational data work from the hacking/computer science side and (3) serious statistical model and inference analysis. Business leaders need to know that all three area are of critical importance and analysts in the data science space need to be experts in all three areas.

[1] 100% was his response to a question – how much of this thought process was emphasized by your education as an economist? Economists are encouraged to start problem solving with storytelling unencumbered by the actual data limitations and unbiased by the influence of a hasty and premature running of a PROC CORR or PROC REG.

CONCENTRATION ON CROSS SECTIONAL PROBLEMS, DATA AND STATISTICS

Cross section data is data that is structured into a table of rows of observations and columns of variables. Typically, there is no time dimension meaning that all of the observations are measured at the same point in time. Such as shown in Figure 1.

Figure 1: Cross Section Example

observation	year	employed	age	Last wage
N_1	2018	Yes	25	12.50
N_2	2018	Yes	18	10.47
N_3	2018	No	19	7.75
N_4	2018	Yes	31	23.90
…	…	…	…	…
N_n	2018	No	33	25.50

When you have multiple cross sections with similar or the same variables, but measured at least during two points in time you have either panel data or longitudinal data. Panel data will contain the same variables, but the observations in different time periods are not from the same sample. A panel data could arise when survey questions have been repeated over time, but the sample respondents in each time period are different.

Longitudinal data follows the same persons or observations over time. Each person is interviewed in each time period. So a question “do you plan to graduate?” asked of freshmen can be compared 4 years later with “Did you graduate?” answered by the same person. These possibilities are shown in Figure 2.

Figure 2: Panel or Longitudinal Data

observation	year	employed	age	Last wage
N_1	2013	Yes	20	7.00
N_2	2013	No	13	.
N_3	2013	No		
N_4	2013	Yes		
…	…	…		
N_n	2013	yes		

observation	year	employed	age	Last wage
N_1	2018	Yes	25	12.50
N_2	2018	Yes	18	10.47
N_3	2018	No	19	7.75
N_4	2018	Yes	31	23.90
…	…	…	…	…
N_n	2018	No	33	25.50

For our illustrations in this paper we focus on a typical one period cross section.

EXPLORATION IS ONLY ONE POINT IN THE APPLIED ANALYTICAL PROCESS.

Theoretical statistics and econometrics is mostly about using techniques to get best[2] estimates and to properly apply the mechanisms of testing. To do applied analysis one must go well beyond just the estimations and inference. Good applied work begins with problem articulation, continues though acquiring and cleaning data and finishes with sophisticated model specification considerations. If we have just a vague (or no idea) what we are looking for, assume that the data is both appropriate for analysis and clean, and assume that our model is known and pre-ordained then we have experienced three points of massive failure in Applied Analysis.

Steps of Applied Analysis

1. Problem Articulation
2. Data Cleaning including exploratory data analysis
3. Model Selection and Specification
4. Reporting of results

PROBLEM ARTICULATION

The first step in any analysis is to ask the right question. Even when the proper question is asked, the meaning might be skewed by any number of confounding issues. The proper articulation of the problem should lead directly to any number of testable hypothesis.

In this paper, we express the problem as do men and women on a college campus pay the same for their living arrangements? Our null hypothesis is then that men and women pay the same to be rejected if possible in the favor of either women pay more or men pay more.

Like the illustration of the data scientist above, good problem articulation requires that we understand the problem theoretically and that we can tell the story of how the price of apartments arise. In this case a story be woven of why rents would be different on the basis of the sex of the renter, given we understand how rents are determined at all. One could jump to the conclusion that any difference would be discriminatory, but discrimination requires that we compare people and their choices exactly and apartments are clearly not all equal and neither are the renters.

Obviously this is a problem of supply of apartments and the demand for those apartments, but what influences both supply and demand? Let me suggest that the story may be that rents should be higher the larger the size of the apartment, the more preferred the location, the higher the safety of the neighborhood, the higher the quality of the living quarters, and the higher the price of non-apartment substitutes such as dorms and houses. Data I might want also include square footage, crime stats, is the apartment on a bike path, price of dorm rooms, and more. We even need to know whether utilities are included and if it is furnished. Unless we generate our own survey these data may not be available. In our case we have a given data set collected by others and we cannot supplement it. It is beyond the scope of this paper to look at the measurable effects of not having quality or size of the apartments, but just because we don't have them does not mean the theoretical effects are not present.

The remainder of the paper concentrates on EDA and data cleaning and touches only a bit on Model Selection and Specification, and very little on the reporting of result.

THE DATA FOR OUR EXAMPLE

The data are a sample of 32 persons renting apartments on the Ann Arbor campus of the University of Michigan as shown in Table 1Table 1: The Ann Arbor Rental Data.[3] Data on quantity is available in number of rooms, but not square feet, and data on location is measured in distance from the campus, but nothing about the quality and safety of the neighborhood those blocks of distance take you through.

[2] Best in the statistical sense is that estimator that is characterized by minimum variance and is unbiased. A special form is said to be BLUE (best linear unbiased estimator) when the estimator is known to be the best among the class of all linear estimators.

[3] The data comes from an example in the text by Pyndick and Rubenfeld (1997).

There is no data on the inclusions of utilities or furniture or competing prices. In the latter case we can assume with some risk to out analysis that each of the 32 rents sampled are all based on a prevailing set of practices such as all of the rents include utilities or none do. Using this sample means we have to be careful in our interpretation of results and as always we should express our results with humility. The data cannot reveal truth, but it can certainly allow us to examine our inferences. It is beyond this paper to go further into model specification and the effect of missing variables, but just ask if you want more.

Here are the data with their explanation.

Table 1: The Ann Arbor Rental Data

Variable	Explanation
RENT	Monthly apartment rent (dollars)
NO	Number of persons living in the apartment
RM	Number of rooms in the apartment
SEX	Sex
DIST	Distance in blocks from campus
RPP	Rent per person (RENT/NO)
RMPP	Rooms per person (RM/NO)

We may think that the 5 original variables are of good quality and clean within reason, but we are still obligated to test that assumption. We can use basic exploratory data analysis[4] to look at individual observations that might be dirty or otherwise influential.[5] Cleaning data (including managing, wrangling and exploring) is the critical next step.

A NOTE ON SEX

The variable SEX is a dummy variable given in the data set with that name with values 0 and 1 that correspond to males (if sex=0) and females (if sex=1). A good practice is to rename or modify an ambiguous variable like this into a readable variable such as FEMALE, a dummy variable with values 1 (if sex=1) and 0 (if sex=0). To rename SEX to FEMALE use:

```
RENAME Sex=Female;
```

To create a dummy variable separately from the old variable use:

```
If SEX=1 then Female=1; Else Female=0;
```

Alternatively you can use:

```
LABEL SEX='Female (1=female, 0=male)';
```

Then within many procedures the title of the variable, SEX, is replaced, by the label.

[4] Kennedy (2008) warns of the dangers of EDA and that warning is mostly based on the understanding that truth is not in the data and that analysts will be led astray by seeing correlations, for example, without having first thought seriously about the problem. This is accomplished in this case by the problem articulation stage conducted before exposure to the data and is exactly the reason our quoted data scientist conducts a thought process with his team before allowing them to discover relationships in the data.

[5] Request my paper "Don't let influential data observations kill your regression and your career," paper prepared for presentation to SESUG, October 2019. Much of this paper depends on looking at influence statistics to seek outliers after the analysis. This paper was withdrawn from SESUG, but will be available soon on request.

A format can also be assigned as shown below and then associated with the variable, SEX:

```
proc format;
value gender      1 ='Female'
0 ='Male';
run;
```

All three of four of these lead to readable output and do not make the reader have to find a code book to define the SEX variable or to seek guidance on how to interpret the results.

HASTY REGRESSION

"I was so excited to get our data, the first thing I wanted to do was run a regression," a well published friend and careful data user said this to me recently in a voice that mixed both the excitement of new data and the tone of guilt because she knew it to be "wrong."

My experience with students and with professional researchers is the tendency to rush to a favorite analysis on a data set. Say, I just acquired a data set and I am anxious to run a model that appeals to me, perhaps just any model, just to see what happens. I load the data into a SAS data set and reach for my PROC REG or other favorite procedure and I am just a moment away from results. Indeed I may use:

```
PROC REG;
 Model RPP = RMPP DIST SEX;
 Run;
```

I run the program and I find that Females pay $20.22 more than the males in the sample holding RMPP and DIST constant, however, the effect of SEX on rent is not statistically significant from zero. Shall I stop investigation here and report that there is no difference in rents based on gender? This may be the case if there is too much pressure from business leaders to get results quickly. As stated above: Data takes time.

As to the result, the hasty wage gap is nonexistent, should I do more work? Do I have confirmation bias such that I suspect subconsciously that there really isn't an effect? If so I may conclude it is good enough. However, the results may be the result of dirty data. Have I even asked the right question?

Did I articulate the problem in the most complete way and in a way that allows for the appropriate testing of the effect? Asking the right question is step 1. Even if you look only at the test of SEX in a model, it is important to know that each of these models yield that test, but they are all different.

A statistical test is not independent of the variables included in the model and as important the variables left out. This is shown dramatically below:

```
PROC REG;
 Modela: model RPP = SEX;
 Modelb: model RPP = RMPP SEX;
 Modelc: model RPP = RMPP DIST SEX;
 run;
```

With the results:

Model 1a – Females pay $37.81 more (p-value=0.033)

Model 1b – Females pay $34.92 more (p-value=0.060)

Model 1c – Females pay $20.00 more (p=value=0.238)

The first step is an unadjusted gap, the second is adjusted or controlled by space per person, and the third is the second, also controlling for the distance of the apartment from the center of campus. All test the same null and alternative hypothesis, but they differ on the maintained hypotheses, e.g., the variables held constant (and in the model).

The third step of good applied analytics is proper model specification and is dependent on (step 1) good problem articulation (the correct question having been asked) and on (step 2) clean data (the proper work having been done to clean the data).

This is why subjecting the data to basic and descriptive statistics is so important. In short we need to look at the data before we do our statistical modeling (this paper) and after we get our result (see paper referred to in footnote 5).

EXPLORATORY DATA ANALYSIS

This paper highlights many methods to examine the data. We are looking to characterize the data, determine its distribution and to understand the data generating process (DGP). Additionally, we are looking for values of the data that may be extreme, or seem not to fit the remaining data. These are often identified as outliers, data points that lie so far away from the bulk of the data that they represent a mistake or, if not a mistake, may represent an individual very different from the bulk of the data.

BASIC DISCRIPTIVE STATISTICS

The most basic and known descriptive measure is the measures of central tendency, which is the average (mean) or most observed value (mode). This is the center of the data and we can ask what the mean rent per person is for women ($166.17) and men ($126.40). We combine that with asking how do the data points spread or disperse around the mean, the standard deviation. This value is the square root of the variance and has the benefit of being in the same units as the mean. The standard deviation of rent per person is $36.18 for women and $47.41 for men. So the mean measure for men is lower than women, but the spread around the mean is larger for men than women as well.

These statistics are easily found with PROC MEANS.

PROC MEANS

Since our question concerns gender differences it is first useful to examine how the variables of interest are affected by the class (or classification) variable sex. We find the mean and standard deviation by PROC MEANS. Maxdec=2 puts the number of decimals printed at 2 and the format statement (explained above) replaces the 0 and 1 codes of the SEX variable in defined words that make the output shown as Table 2 more readable:

```
proc means data=rentdata.rent2 maxdec=2;
     class sex;
     var rpp;
     format sex gender.;
     run;
```

Table 2: PROC MEANS default output

Analysis Variable : RPP rent per person = RENT/NO						
Female(1=female, 0=male)	N Obs	N	Mean	Std Dev	Minimum	Maximum
Male	22	22	126.35	47.41	50.75	285.00
Female	10	10	164.17	36.18	115.00	245.00

Default statistics include the sample size, N, and the minimum and maximum values of the data for each breakdown of the class variable. It is interesting that the men in the sample pay rent per person as low at $51 and as high as $285. This should suggest that there is something going on in the data if men are willing to live cheaper and also at the same time more expensively than females. Typically that "something going on" is known as heterogeneity of the data and suggests a difference in the data generating process (DGP)

In the next procedure we add 3 other statistics: CV (coefficient of variation), skewness, and kurtosis while dropping min and max for space. In this case the only statistics to be printed are those listed as options in the PROC statement:

```
proc means data=rents maxdec=2 n mean std cv skew kurtosis;
    class sex;
    var rpp ;
    format sex gender.;
    run;
```

The output is shown in Table 3.

Table 3: PROC MEANS output with selected statistics

Analysis Variable : RPP rent per person = RENT/NO							
Female(1=female, 0=male)	N Obs	N	Mean	Std Dev	Coeff of Variation	Skewness	Kurtosis
Male	22	22	126.35	47.41	37.52	1.70	5.16
Female	10	10	164.17	36.18	22.04	1.17	2.03

The coefficient of variation statistic is defined as the ratio of the standard deviation to the mean and is presented in percentage terms. It is also known as the relative standard deviation, showing what percentage of the mean is the standard deviation. Larger numbers indicate less precise means and less stability of the data. The CV is a better measure to make the point suggested in the last paragraph that something is going on since men will live both cheaper and in more expensive arrangement.

Skewness measures whether the distribution of the variable is larger in one direction or the other. A normally distributed variable would have skewness equal to zero and by comparison, neither female nor male RPP distributions are likely to be normal and the positive skew is far more apparent in the male sample than the female sample. Kurtosis which measures heaviness of tails of the distribution would be zero for a normal distribution and shows males to have much more mass in the tails of the RPP distribution.

CONFOUNDING VARIABLES

The other variables in the data set confound the relationship between RPP and SEX and are factors to be held constant. Again PROC MEANS is run on the variables RMPP and DIST by SEX and the results shown in Table 4.

Table 4: PROC MEANS showing statistics on confounding variables

Female(1=female, 0=male)	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Male	22	rmpp	Rooms per person	22	0.9	0.2	0.5	1.0
		DIST	distance from center of campus in blocks	22	15.7	14.8	3.0	60.0
Female	10	rmpp	Rooms per person	10	1.0	0.1	0.7	1.0
		DIST	distance from center of campus in blocks	10	7.4	6.4	0.0	24.0

We see that women are likely to live closer to campus than are men and the RMPP variable looks to be about the same. What PROC MEANS fails to show is any kind of visual representation of the data and we turn our attention to that now.

So far we know women pay higher rents on average, have less deviation in those rents and live closer to campus. But individual observations could be obscuring more information or influencing the results.

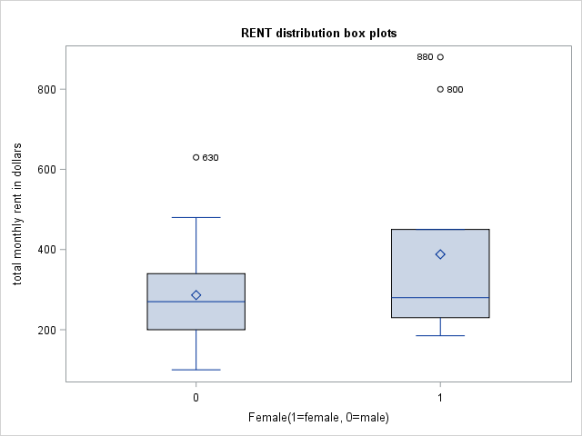
VISUALIZATION OF YOUR DATA

PROC SGPLOT is extremely useful and gives a number of choices for graphing data. The first example produces a vertical box plot of the RENT data separately by the categorical variable SEX:

```
proc sgplot data=rents;
 vbox rent / category = sex datalabel;
 run;
```

VBOX shows the boxes vertically and HBOX would show them horizontally. The choice is mostly a matter of preference. This box plot produces a plot shown in Figure 3 that shows at a glance the distribution of points marking means, modes, quartiles and so much more. Especially poignant is the outliers observed in the tails. The datalabel option prints out the value of each outlier. In the rent variable by SEX we see three rather large RPP outliers, $630 for females and $800 and $880 for females. These can have an undue influence on the analysis and we will look at that below.

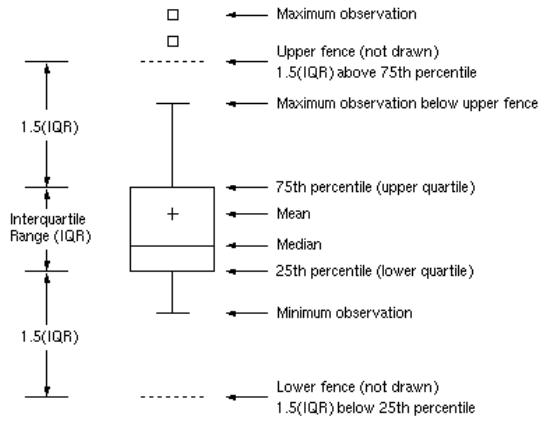
Figure 3: Box plot showing extreme values for RENT



This chart in Figure 4 from the SAS documentation tells how to read a common box plot.[6]

[6] SAS 9.4. SAS/STAT User's Guide. The BOXPLOT Procedure. Accessed at https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=statug&docsetTarget=statug_boxplot_details09.htm&locale=en

Figure 4: How to read a box plot

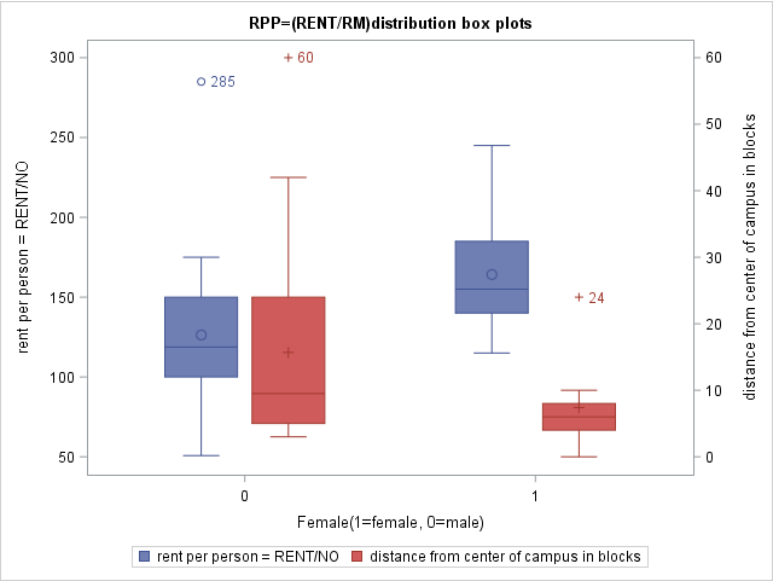


The code to display the box plots is shown below. Two variables are shown overlaid in the next set of code and focus on the more relevant variable RPP and the DIST measure. Setting boxwidth is a manner of taste (the default is the program chooses) and discreteoffset is used so the two plots are not printed on top of each other. Also notice the DIST variable is measured in blocks and not in dollars so that graph is assigned to the right hand vertical axis by the use of the yaxis2 option:

```
Title1 'RPP=(RENT/RM)distribution box plots';
proc sgplot data=rents;
     vbox rpp  / category = sex  boxwidth=0.25 discreteoffset=-0.15 datalabel;
     vbox dist / category = sex  boxwidth=0.25 discreteoffset=+0.15 datalabel y2axis;
 run;
```

The above code produces the comparative box plots as shown in Figure 5. Comparing the distributions of RPP show higher measures of central tendency and greater range for females compared to males, and a large outlier for males as one renter pays considerably higher rents than all the others exceeding by far the 75th percentile. In the case of DIST we see visually what we could discern in the PROC MEANS that males have a much greater range of data while females prefer to live close to campus. What we could not easily see in the PROC MEANS data is the MAX value for each gender is indeed an outlier.

Figure 5: Box Plots for RPP and DIST by SEX



PROC UNIVARIATE AND EXTREME OBSERVATIONS

On the outliers shown, their value is shown in the box plot, but PROC UNIVARIATE offers another look at our data and by default identifies the top and bottom (extreme) five values

Rather than allow the pages of results of PROC UNIVARIATE, ods select is set to only pull the extreme values and observation numbers.[7] The code here is:

```
ods select extremeobs ;
proc univariate data=rents;
 var rpp dist;
 class sex;
 run;
```

This code produces four tables, two (by SEX) for RPP and two (by SEX for DIST). To save space only two tables are shown in Table 5: Univariate Extreme Values for RPP by SEX and you can see the value and the observation number for the lowest and highest values. We already saw that RPP=$285 for males was an extreme value in the box plot, and now we know it is observation number 4. Curiously, RPP=$245 among the females seems like it could be extreme, but was not highlighted on the box plot.

Table 5: Univariate Extreme Values for RPP by SEX

The UNIVARIATE Procedure
Variable: RPP (rent per person = RENT/NO)
SEX = 0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
50.75	17	155	8
80.00	14	155	12
85.00	2	160	19
87.50	3	175	15
92.50	9	285	4

The UNIVARIATE Procedure
Variable: RPP (rent per person = RENT/NO)
SEX = 1

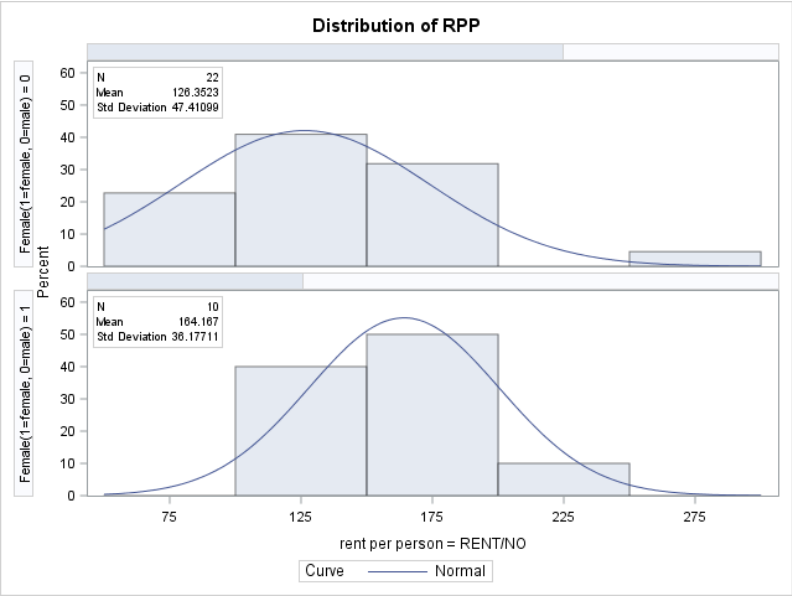
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
115.00	29	160	31
140.00	27	170	24
140.00	26	185	30
146.67	28	190	23
150.00	25	245	32

PROC UNIVARIATE can also display histograms of the distribution of our variable of interest as shown in Table 6: Histograms and Normal Distribution for RPP by SEX. The following code creates the desired output. (note that PROG SGPLOT can also generate the histograms):

```
proc univariate data=rents noprint;
     class sex;
     histogram RPP / midpercents nrows=2 intertile=1 cprop normal(noprint);
     inset n = "N" mean std / pos = nw;
     run;
```

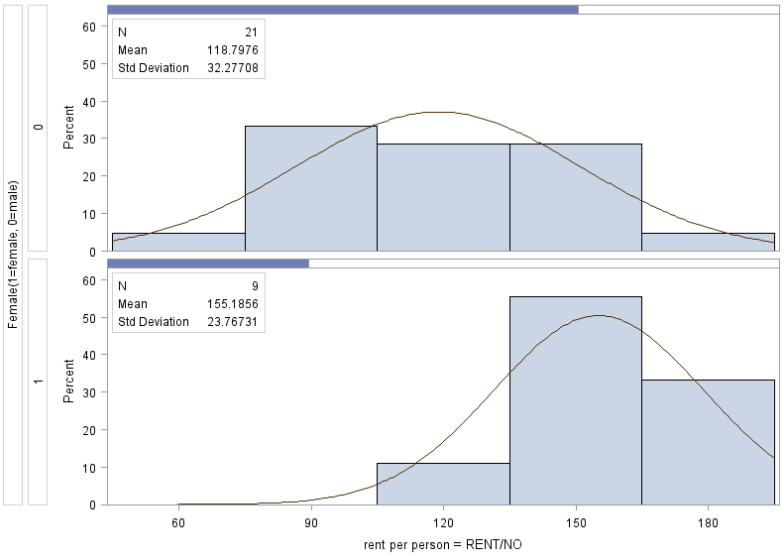
[7] Here is a fast way to find the name of ODS tables: (1) Go to the first Appendix of the *ODS User's Guide* (2) Select the product and (3) Select the procedure, Wicklin (2015)

Table 6: Histograms and Normal Distribution for RPP by SEX



In our first demonstration of the power of outliers, the histograms are rerun with the two largest outliers ($245 for females and $285 for males) are removed and the visual results are dramatic as can be seen in Figure 6. Also note the dramatic effect on the mean and standard deviation.

Figure 6: Histograms and Normal Distributions for RPP and SEX after removing two extreme values.



ALTERNATIVE TO DELETING OBSERVATIONS IN MULTIVARIATE ANALYSIS

It is not always an easy thing to decide whether to delete the observation of the outlier, to set it to missing, or to “mark it” by creating a dummy variable which is equal to 1 when the observation is the outlier and zero for all other observations. Using a separate mark for each outlier can have a useful effect in multivariate analysis. For example, the regression parameter estimates of the following two models are identical. I leave that for you to demonstrate. Using the M1=1 mark in this way preserves a valuable observation rather than deleting the entire record:

```
data rent2;
 set rents;
 if RPP=285 then M1=1; else M1=0;
 run;
Proc reg data=rent2;
 model RPP = RmPP DIST SEX M1; /* N=32 */
 run;
Proc reg data=rent2;
 model RPP = RmPP DIST SEX ; /* N=31 */
 where m1=0;
 run;
```

The two regressions produce identical parameter estimates as can be seen in the regression output at the end of the paper when more exceptions are marked.

PROC FREQ

Another way of exploring the data makes use of PROC FREQ. The tables statement in the syntax asks for two separate 2 way tables. The first table is RMPP*SEX and is shown below followed by the second table which is RM*NO and gets to the definition of RMPP:

```
Title1 'PROC FREQ distance by sex';
proc freq data=rents;
     tables rmpp*sex rm*no
            /norow nocol nocum noprecent;
     run;
```

Table 7: PROC FREQ two way table of RMPP and SEX

PROC FREQ distance by sex

The FREQ Procedure

Frequency	Table of rmpp by SEX			
		SEX(Female(1=female, 0=male))		
	rmpp(Rooms per person)	0	1	Total
	0.5	5	0	5
	0.6666666667	1	1	2
	1	16	9	25
	Total	22	10	32

In Table 7 we discover something that none of the other procedures above reveal, 30 of the 32 renters apparently either share a room with another (RMPP=0.5) or have a room to themselves. And only 2 observations are RMPP=0.67) and seemingly unique.

The second table of RM*NO gets at the definition of RMPP. We discover again that 25 are in apartments with one room per person. The other room combinations are 5 persons in combinations of 2 in one room and 4 in 2 rooms. And 2 renters are in apartments of 2 persons in three rooms.

By exploring the derivation of RMPP we may learn something that affects how we will model our problem. It is important that we understand the data generating process, for example just by knowing that RMPP = 1 we might wrongly assume everyone is living in the same type of apartment, where 1 person in 1 room is potentially the same as 6 in 6 rooms.

Table 8: PROC FREQ two way table of RM by NO

Frequency	Table of RM by NO							
		NO(number of persons in apartment)						
	RM(number of rooms)	1	2	3	4	5	6	Total
	1	6	3	0	0	0	0	9
	2	0	13	2	2	0	0	17
	3	0	0	3	0	0	0	3
	5	0	0	0	0	1	0	1
	6	0	0	0	0	0	2	2
	Total	6	16	5	2	1	2	32

USE OF PROC FORMAT IN PROC FREQ, MEANS AND OTHER PROCEDURES

What more can we learn? The example of Table 9 uses PROC FORMAT to display the data in categories. This prevents the frequency distribution from listing each value on a separate row or column. In this case, the values statement defines the values that will be in each category. Three DIST categories allows us to see the distribution by being close, further out and far out from campus center. Clearly we can see that females do favor close and more males prefer far out by running the following code:

```
proc format data=rents;
 value distance  0-6    ='close 0-6 blocks'
                        7-12   ='further 7-12 blocks'
                        13-60  ='far out 16-60 blocks';
value gender            1      ='Female'
                        0      ='Male';
    run;

Title2 'PROC FREQ distance categories by sex';
proc freq data=rents;
     tables dist*sex /norow nocol nocum noprecent;
     format dist distance. sex gender.;
     run;
```

Table 9: PROC FREQ using format statement

PROC FREQ distance categories by sex

The FREQ Procedure

Frequency	Table of DIST by SEX			
	DIST(distance from center of campus in blocks)	SEX(Female(1=female, 0=male))		
		Male	Female	Total
	close 0-6 blocks	7	6	13
	further 7-12 blocks	7	3	10
	far out 16-60 blocks	8	1	9
	Total	22	10	32

Not everything that we may look for will prove important, but we are learning about the data each time we try. Knowing the data generating process is critical for cleaning data and for good data modeling.

PROC MEANS can also be used with the format statement to calculate statistics on the resulting groups cells as the next example shows:

```
Title2 'PROC MEANS distance categories by sex';
proc means data=rents maxdec=2;
     class dist sex;
     var rent rpp  ;
     format dist distance. sex gender.;
     run;
```

The results are in Table 10

Table 10: PROC MEANS using format statements for DIST and SEX

PROC MEANS distance categories by sex

The MEANS Procedure

distance from center of campus in blocks	Female(1=female, 0=male)	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
close 0-6 blocks	Male	7	RENT	total monthly rent in dollars	7	275.00	146.46	100.00	450.00
			RPP	rent per person = RENT/NO	7	140.36	69.32	85.00	285.00
	Female	6	RENT	total monthly rent in dollars	6	403.33	248.73	190.00	880.00
			RPP	rent per person = RENT/NO	6	156.11	19.93	140.00	190.00
further 7-12 blocks	Male	7	RENT	total monthly rent in dollars	7	260.00	63.05	160.00	310.00
			RPP	rent per person = RENT/NO	7	122.86	31.90	80.00	155.00
	Female	3	RENT	total monthly rent in dollars	3	405.00	342.82	185.00	800.00
			RPP	rent per person = RENT/NO	3	153.33	35.47	115.00	185.00
far out 16-60 blocks	Male	8	RENT	total monthly rent in dollars	8	319.50	158.03	200.00	630.00
			RPP	rent per person = RENT/NO	8	117.16	38.05	50.75	175.00
	Female	1	RENT	total monthly rent in dollars	1	245.00	.	245.00	245.00
			RPP	rent per person = RENT/NO	1	245.00	.	245.00	245.00

USING CUSTOM TABLES WITH PROC TABULATE

Of course, not all investigations are only exploratory and often begin to reveal useful tables for your reporting. PROC TABULATE is extremely useful because of how it lends itself to customization. All variables to be included in the table produced by PROC TABULATE must be listed in either the class or var statement. In the Table statement the comma separates the variable syntax into what is on the rows and what is on the columns:

```
Title1 'Custom Tables';
Title2 'Two dimensional table, rows before the comma, columns after.';
Title3 'Sex and RPP interacted with one stat requested.';
Proc tabulate data=rents;
     class dist no rm sex ;
     var rent rpp  ;
     Table dist , sex*rpp*mean;
     format dist distance. sex gender.;
     run;
```

The results are shown in Table 11.

Table 11: PROC TABULATE showing means of RPP by DIST and SEX

Custom Tables for RPP Statistics
Two dimensional table, rows before the comma, columns after.
Means of RPP by Sex and distance.

	Female(1=female, 0=male)	
	Male	Female
	rent per person = RENT/NO	rent per person = RENT/NO
	Mean	Mean
distance from center of campus in blocks		
close 0-6 blocks	140.36	156.11
further 7-12 blocks	122.86	153.33
far out 16-60 blocks	117.16	245.00

We can add a total column and add sample sizes and standard deviations with the following code which produces Table 12 from the following code:

```
Title2 'Two dimensional table, rows before the comma, columns after.';
Title3 'Sex and RPP interacted with three statistics requested.';
Title4 'Use of titles and null titles to clean up the presentation';
Proc tabulate data=rents;
     class dist no rm sex ;
     var rent rpp  ;
     Table (all=Total dist) , (all='Total' sex='')*rpp=''*(n mean stddev);
     format dist distance. sex gender.;
     run;
```

Table 12: PROC TABULATE with multiple statistics and formats applied.

Custom Tables for RPP Statistics
Two dimensional table, rows before the comma, columns after.
Mean, standard deviation and sample size of RPP by Sex and distance.
Use of titles and null titles to clean up the presentation

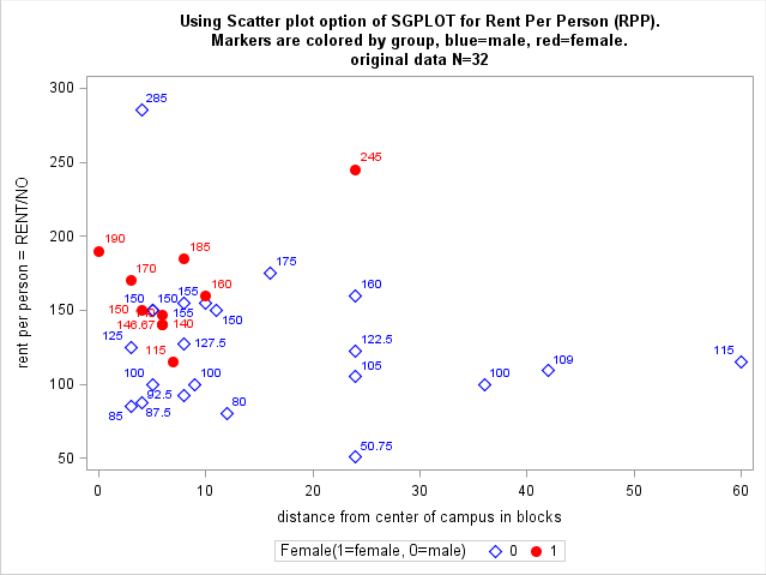
	Total			Male			Female		
	N	Mean	StdDev	N	Mean	StdDev	N	Mean	StdDev
TOTAL	32	138.17	47.11	22	126.35	47.41	10	164.17	36.18
distance from center of campus in blocks									
close 0-6 blocks	13	147.63	51.33	7	140.36	69.32	6	156.11	19.93
further 7-12 blocks	10	132.00	34.27	7	122.86	31.90	3	153.33	35.47
far out 16-60 blocks	9	131.36	55.52	8	117.16	38.05	1	245.00	.

SCATTERPLOT

Another graphical approach also uses SGPLOT with the scatter plot option. In this case the styleattrs (style attribute) command sets the symbols and colors while the size of the markers are set in the option of the scatter command. The results are shown in Figure 7 which is generated by the following code:

```
ODS GRAPHICS / ATTRPRIORITY=NONE;
title2 'Using Scatter plot option of SGPLOT for Rent Per Person (RPP).';
title3 'Markers are colored by group, blue=male, red=female.';
title4 'original data N=32';
proc sgplot data=rent2;
    styleattrs datasymbols=(diamond circleFilled ) datacontrastcolors=(blue red);
    scatter y=RPP x=dist  / group=sex  Markerattrs=(size=10px)datalabel ;
    run;
```

Figure 7: Visualization of RPP by SEX using SGPLOT scatter command



ADVANCED STATISTICS – STATISTICAL INFERENCE

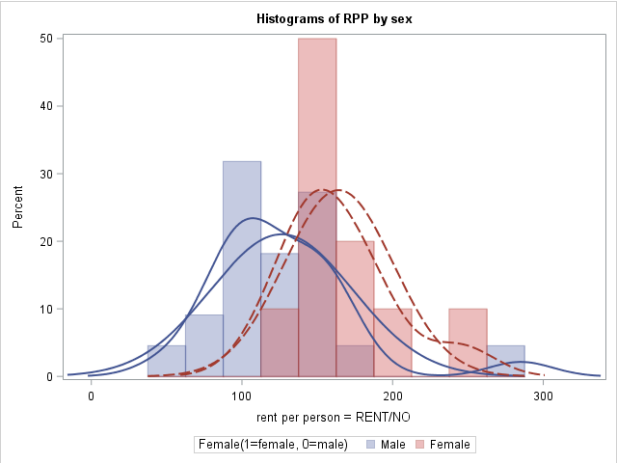
While we have explored much about our data set, we return to the overall question and bring statistical procedures to bear on the ultimate question.

The first is to return to the PROC SGPLOT and map the distributions of the RPP by SEX and to determine how close to unimodal and normal they are before applying statistical tests. In the following SGPLOT code we use the histogram command and plot normal and kernel densities implied by the data:

```
Title1 'Histograms of RPP by sex';
proc sgplot data=rents;
     histogram RPP / Group=sex transparency=.6 binwidth=25;
     density RPP / group=sex type=kernel;
     density RPP / group=sex type=normal;
     format dist distance. sex gender.;
     run;
```

The results are shown in Figure 8.

Figure 8: PROC SGPLOT histogram of RPP by SEX

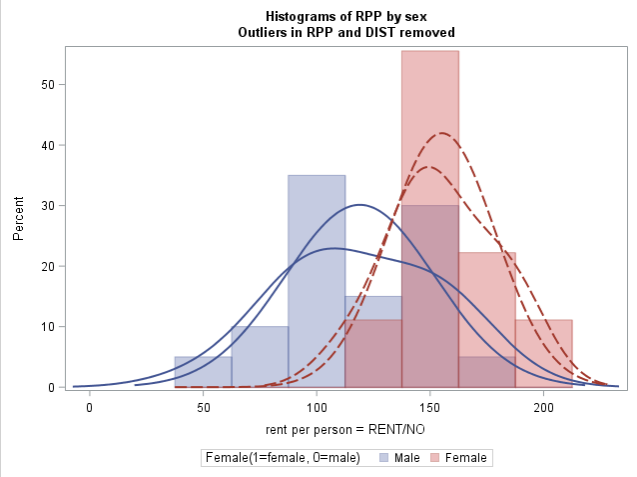


Above we identified at least 3 outliers and the code in the data step below identifies and marks each outlier. The variable clean is created to be yes for all observations that are not outliers in the following DATA STEP:

```
data rent2;
 set rents;
 length clean $3;
 if sex=0 and RPP=285 then M1=1; else M1=0;
 if sex=0 and dist=60 then M2=1; else M2=0;
 if sex=1 and dist=24 then M3=1; else M3=0;
 if sum(of M1, M2, M3) = 0 then clean='yes'; else clean='no';
 run;
```

Once we rerun the histogram on the clean data set (using where clean=’yes’) we see in Figure 9 that both distributions look more approximately normal. We may want to experiment with other outliers, but for purpose of this paper, let us proceed.

Figure 9: PROC SGPLOT histograms of RPP by SEX on Cleaned data.



LINEAR VERSUS NONPARAMETIC REGRESSION OF RPP ON DIST AND SEX.

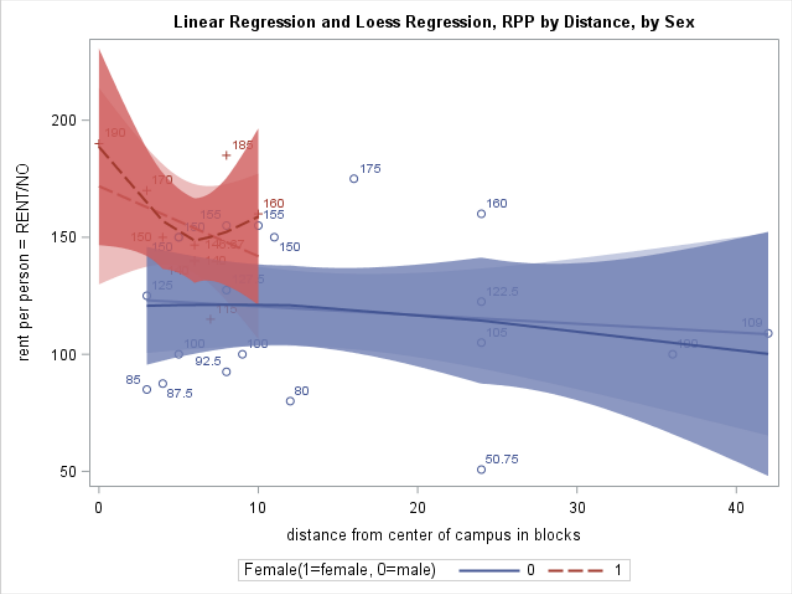
PROC SGPLOT allows us to visualize the fit of RPP on DIST separately by SEX. The reg line in SGPLOT gives the predictive results. The predictive analytics here are restricted to be linear and you can see from the graph the regression is shown with a straight line, one for females and one for males. The 95% prediction interval is shown and shaded around the linear regression.

On the same graph we overlay a LOESS regression. This is a near neighbor weighted non-parametric regression. It is useful when you do not know a suitable parametric regression and offers a perfect contrast to the linear regression specified by the reg statement. The process iterates over all observations in k clusters of observations at a time. In each cluster a regression is fit giving more weight to the closest observation and low to zero weight to observations further away. This is repeated for all clusters of observations. Furthermore, the LOESS procedure is suitable when there are outliers in the data and a robust fitting method is necessary.[8] The code asks for the prediction line to be drawn with error bands (clm) and to include no markers. Transparencies are chosen to make it possible to see both lines and error bands when overlaid on the graph.

```
Title2 'Linear Regression and Loess Regression, RPP by Distance, by Sex';
proc SGPLOT data=rent2;
     reg x=dist y=rpp /group=sex clm clmtransparency=.6 datalabel;
     loess x=dist y=rpp /group=sex clm nomarkers clmtransparency=.2;
     where clean='yes';
run;
```

The results of Figure 10 are quite dramatic showing visibly the greater distance that males live from campus, with the LOESS regression very similar to the linear. For females the results of the LOESS regression is revealing and dramatic and decidedly nonlinear. Outliers observed are now observed AFTER the linear regression results and strictly are those who are outside the 95% confidence interval for the linear regression. Not only do we see that men and women are quite different, we see that women are different from other women. Perhaps there is two separate preference sets for women, which is a way of saying that women aren't women (women aren't a homogeneous group) in their distance preferences.

Figure 10: PROC SGPLOT Linear regression and LOESS regression



[8] See Overview of LOESS procedure at https://documentation.sas.com/?docsetId=statug&docsetVersion=15.1&docsetTarget=statug_loess_overview.htm&locale=en/

As to the outliers, these are not the result of data we knew to be dirty beforehand, but rather in the words of Larry Klein, represent an indication or measure of our ignorance in applying the specific linear and nonlinear models to our data. This is the subject of the paper referenced in footnote 5.

PROC CORR TO SHOW PAIRWISE RELATIONSHIPS

Calculating pairwise correlations on our data will often reveal relationships and allow them to be tested against a null hypothesis of no or zero correlation. Correlations can be positive or negative and are often one of the first things a researcher is likely to do. One can learn much here, but only whether there exists a strong and linear relationship between the Y and X variable. There is a one-to-one correspondence between a zero order correlation of X and Y, and a 2 parameter regression of Y on a single X. The code is:

```
Title2 'PROC CORR on full data set';
proc corr data=rent2;
     var rpp; with rmpp dist sex;
     run;
```

Results are in Table 13.

Table 13: PROC CORR on the full data set, Pearson zero-order correlations

Pearson Correlation Coefficients, N = 32 Prob > |r| under H0: Rho=0	RPP
rmpp Rooms per person	0.49457 0.0040
DIST distance from center of campus in blocks	-0.20103 0.2699
SEX Female(1=female, 0=male)	0.37797 0.0329

The effect of the three original outliers can be seen in Table 14 by restricting PROC CORR to only the clean observations by use of the where variable from the following code:

```
Title2 'PROC CORR on clean data set';
proc corr data=rent2;
     var rpp; with rmpp dist sex;
     where clean='yes';
     run;
```

Table 14: PROC CORR on the clean data set, Pearson zero-order correlations

Pearson Correlation Coefficients, N = 29 Prob > |r| under H0: Rho=0	RPP
rmpp Rooms per person	0.58963 0.0008
DIST distance from center of campus in blocks	-0.30390 0.1090
SEX Female(1=female, 0=male)	0.49292 0.0066

Let's compare the correlations from the original and clean data set.

As we can see the correlations of RPP with RMPP is 0.49 on the full data and rises to 0.60 on the clean data set although both have p-values sufficiently low to reject the null hypothesis of no relationship. In the case of RPP with SEX the coefficient rises from 0.38 on the full data to 0.49 on the clean data set and both have low or somewhat low p-values (although the clean-data correlation would not be significant at a 5-perent level.

Finally the correlation of RPP with DIST is negative and is stronger in the clean data set although neither reach significance at a 10 percent level. So cleaning the data strengthened in this case all of the pairwise relationships. However, the issue is that uncontrolled pairwise relationships are not sufficient to have an answer here. We might be tempted to ignore DIST on the basis of the correlations, but our eyes have already convinced us that it has a real differentiating effect.

So PROC CORR allows us to look at these pairwise relationships holding other variables constant through the use of partial correlations. In the example below we look at the correlation between RPP and SEX holding constant the other variables of DIST and RMPP:

```
Title2 'PROC CORRELATION on clean data set';
title2 'Partial correlation of RPP and SEX, holding constant DIST and RMPP';
proc corr data=rent2;
    var rpp; with sex; partial RMPP DIST;
    where clean='yes';
    run;
```

The results are in Table 15.

Table 15: PROC CORR - partial correlations RPP with SEX, holding constant RMPP and DIST

Pearson Partial Correlation Coefficients, N = 29 Prob > |r| under H0: Partial Rho=0	
	RPP
SEX	0.29695
Female(1=female, 0=male)	0.1325

There are three PROC CORR statements above. The default output is every variable in the var list is correlated by every other variable separately yielding a kxk matrix of results for a var statement with K variables. I have modified each corr to get at something more direct to our problem.

In the first case the entire data set is used and the Correlation between RPP and SEX is .378. In the second the correlation is calculated on the clean data set with the result that the correlation between RPP and SEX rises to .493 with a smaller p-value. Finally, a partial correlation is calculated, that is what is the correlation of RPP and SEX holding constant the values of DIST and RMPP? That answer is a partial correlation of 0.300 with a high p-value of 0.13.

The zero order correlations of .378 and .493 are significantly different from zero, but when holding constant the two variables RMPP and DIST, the significance falls and the null hypothesis that the true value of the relationship between RPP and SEX is zero cannot be rejected at consensual levels of significance. (Notice that we are back where we started with our hasty regression wherein the variable SEX seems to be insignificantly different from zero.)

PROC TTEST A STATISTICAL TEST OF DIFFERENCE BETWEEN TWO MEANS

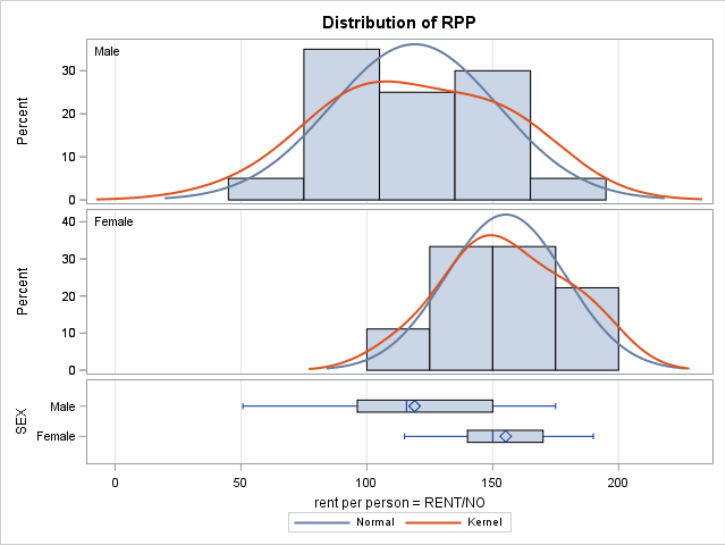
PROC TTEST will reveal much of the same data from PROC MEANS and UNIVARIATE above and focuses on the mean of RPP across the two values of the class variable. It reports the difference in

means and two possible standard deviations of that difference based on whether the variances of the two independent sample are the same or different. Again we run on the clean data set and specify class and variable for the test in the following code:

```
Title2 'Testing the average rent (or RPP) difference between males and females.';
Ods graphics on;
proc ttest data=rent2 ;
     class sex;
     var rpp;
     where clean='yes';
     run;
```

The output of the PROC TTEST is in Table 16 and gives the mean and standard deviations of two independent samples. This procedure calculates the test of the difference in two means, It also shows the results graphically in Figure 11 and we start with that graph (you must have ODS graphics on). By default the histograms that we ran above are automatically created. By inspection of the means in the graph we may suspect that the difference is large and significant.

Figure 11: PROC TTEST ODS distributions of RPP by SEX.



The output of PROC TTEST includes a test of the equality of variances of the RPP variable between men and women. The null hypothesis is that the variances are equal and we can see in the resulting table that the F value is too low to reject the equality of variances (p-value is 0.3397). We therefore rely on the T-test given in the table under the assumption of equal variances which shows that females pay on average $39.20 more than men does reach significance (p-value 0.0066). However, this is a raw or unadjusted gap in rents and does not hold constant any of the variables we care about.

Table 16: PROC TTEST results for RPP(men)-RPP(women)

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	27	-2.94	0.0066
Satterthwaite	Unequal	21.247	-3.34	0.0031

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	19	8	1.94	0.3397

So our conundrum still exists, is there a statistical difference by sex or not. In the final step we turn to a linear regression model, on the clean data and test two hypothesis (1) holding constant DIST and RMPP is RPP statistically different by SEX and (2) Are separate regressions on the basis of SEX statistically different? That is in the latter case if we run separate regressions are they different by SEX.

LINEAR REGRESSION

We have examined the data, made all of the nnecessary adjustments to make the data clean, have determined that womenn do pay more perperson in rents in men, and now seek to ins a model that expresses the data generating process and explains the rents of the persons who are in the sample.

It is time to think back the the original theoretical analysis of the problem and recall that many of the variables we would want are not available. We can assume some things such as all rents rental units identically include or do not include utilities (probably not a bad assumiption) and that the crime on every block is the same (likelly not a good assumption). As we proceed, model specification and estimation would consider much about missing relevant variables and other specification issues, but that takes up beyond the point of this paper.

Given the data limitations and proceeding with humility we offer two possible models, model 1 which assumes the three right hand side (RHS) variables enter linearlly and model 2 where the effect of each RHS variable beyond the SEX variable actually has both a linear and an interactive effect. The code follows, showing models 1 and 2 on the full sample (N-32), models 1a and 2a on the observations in the clean data set (N=29) and models 1b and 2b which include the full data set, but marks the observations that were considered outliers in this case:

```
Data work.rent2;
    Set work.rents;
    If SEX=1 then female=1; elsefemale=0;
    Femrmpp = female*RMPP;
    Femdist = female*DIST;;
    Run;

Title1 'Linear Regression and fully interactive model with and without outlier marks';
proc reg data=rent2;
    model 1:  model rpp = rmpp dist sex;
    model 2:  model rpp = rmpp dist sex femrmpp femdist;
    Test 2:   test sex=femrmpp=femdist=0;
    model 1b: model rpp = rmpp dist sex m1 m2 m3;
    model 2b: model rpp = rmpp dist sex femrmpp femdist m1 m2 m3;
    Test 2B:  test sex=femrmpp=femdist=0;
    run;

Title2 'Linear Regression fully interactive model after outliers removed from data';
proc reg data=rent2;
    model 1a: model rpp = rmpp dist sex;
    model 2a: model rpp = rmpp dist sex femrmpp femdist;
    Test 2a: test sex=femrmpp=femdist=0;
    where clean='yes';
    run;
```

The results of the above code have been transcribed to a publication quality table created in Excel and reporting all of the results of the various models. Table 17 shows those results.

Table 17: PROC REG results of RPP on RMPP DIST and SEX, original and clean data, linear and fully interactive results

(t-statistics in parentheses)

		model 1	model 1a	model 1b	model 2	model 2a	Model 2b
		all data	cleaned	cleaned with marks	all data	cleaned	cleaned with marks
Intercept	**β1**	35.91	48.46	48.46	30.72	41.69 *	41.69 *
		(1.05)	(2.26)	(2.26)	(0.91)	(1.84)	(1.84)
Female	**β2**	20.02	17.79	17.79	99.93	98.27	98.27
		(1.21)	(1.55)	(1.55)	(0.83)	(1.22)	(1.22)
Roomper	**β3**	118 **	97.92 ***	97.92 ***	130.51 ***	105.14 ***	105.14 ***
		(3.02)	(3.93)	(3.93)	(3.35)	(3.97)	(3.97)
Dist	**β4**	-0.79	-0.96 *	-0.96 *	-1.15 **	-0.92 *	-0.92 *
		(-1.35)	(-1.87)	(-1.87)	(-2.)	(-1.75)	(-1.75)
Roomper*Female	**β5**				-121.94	-70.54	-70.54
					(0.96)	(-0.83)	(-0.83)
Dist*Female	**β6**				4.56 **	-2.34	-2.34
					(2.22)	(-0.74)	(-0.74)
M1				122.47			141.86 *
				(5.41) ***			(1.84)
M2				26.34			23.45
				(0.77)			(0.67)
M3				103.93 ***			148.6
				(3.75)			(2.38)
n		32	29	32	32	29	32
MSE		1573	615	615	1411	626	626
F		5.24	9.82 ***	14.49 ***	4.55	6.09	10.86 ***
root MSE		39.66	24.79	24.79	37.56	25.02	25.02
adj R sq		0.29	0.49	0.72	0.36	0.48	0.72

Dependent variable is Rent Per person = RENT per unit / NO of persons per unit.

note: All regressions estimated with OLS SAS REG procedure.

*** significant at the .01 level

** significant at the .05 level

* significant at the .10 level

The effect of being Female in models 1,1a, and 1b is determined by a t-test on the coefficient of Female. Those values are shown in the first three columns and in every case suggest that women and men do not pay different amounts of rent when RMPP and DIST are controlled. From the t-test we draw the conclusion that uncontrolled, there appears to be a significantly higher rent being charged to females, but when controlled the multiple regression specification of model 1 suggest that indeed females to not pay higher rents, the effects of higher rents coming through the other two regressors, RMPP and DIST.

End of story, possibly not. What if the strong showing of RMPP and the weaker association with DIST actually differs between the sexes? Model 2 of Table 17 shows those results.

Table 17 uses a fully interactive model specification.

Let Y = RPP and F=female or the sex dummy variable. Then the model is written as

(1) Y = b1 + b2*F + b3*RMPP + b5*RMPP*F + b4*DIST + b6*DIST*F + e,

such that when F=0 we have

(2) Y = b1 + b3*RMPP + b4*DIST + e

And when F=1 we have

(3) Y = (b1+b2) + (b3+b5)*RMPP + (b4+b6)*DIST + e

So by running equation (1) as shown and labeled as Model 2, 2a and 2b in Table 17 we can see b1, b3, and b4 as the male coefficients in an equation restricted to males, and we can interpret b2, b5, and b6 as the coefficients showing the difference between the male and female coefficients. None of the individual t-tests on b2, b5, and b6 will test the effect of being female on RPP, however all three can be used to represent whether the male and female generation of RPP are the same or not. That is, the null hypothesis that there is no difference between male and female equations is written in SAS as:

```
TEST F=FEMRMPP=FEMDIST=0;
```
[9]

Those tests are shown in Table 18.

Table 18: PROC REG hypotheses tests results using the TEST statement.

Linear Hypotheses	F value		p-value
Model 2			
H0: β2=β5=β6=0	F3,26 =	2.28	0.103
Model 2a and 2b			
H0: β2=β5=β6=0	F3,23 =	1.31	0.296

ABOUT INTERPRETING PARAMETER ESTIMATORS AND TEST RESULTS.

When a continuous variable such as DIST is put in a regression equation with a quantitative dependent variable the resulting estimate such as -0.97 in model 1A, is actually the result of the partial derivative of RPP with respect to DIST, $\partial RPP/\partial DIST$, and each block away from campus reduced average rents per person by about $1.00 per block of distance. We cannot interpret the coefficient of Female the same way because you cannot take the partial derivative of a categorical or dummy variable. So we resort to comparing Expected Value of rents per person. For example, the expected value difference in model 1a would be written as E[RPP | Female=1] - E[RPP | Female=0] = 17.19 in model 1a.

(4) Y = b1 + b2*F + b3*RMPP + b4*DIST + e

(5) E[Y | F=1] - E[Y | F=0] = [b1 + b2*1 + b3*RMPP + b4*DIST +e]
- [b1 + b2*0 + b3*RMPP + b4*DIST +e]
= β2

So in model 1, the differences between RPPP by men and women is β2. We can therefore adopt the null hypothesis, H0: β2 = 0, versus the alternative hypotheses, H1: β2 > 0. This can be accomplished by a

[9] In this case the null hypothesis is Lβ=c and the numerator of the F test is Q=(Lb−c)′(L(X′X)−L′)−1(Lb−c) divided by the number of degrees of freedom, r, (r=3 in this example) where b is the estimate of β. The denominator is s^2. The F then has r and n-k degrees of freedom.

WALD test statement, TEST FEMALE=0; however that test is identical to the t-test on the parameter estimate as found in the default regression output.

(6) $Y = b1 + b2*F + b3*RMPP + b4*DIST + b5*F*RMPP + b6*F*DIST + v$

(7) $E[Y \mid F=1] - E[Y \mid F=0] = [b1 + b2*1 + b3*RMPP + b4*DIST + b5*1*RMPP + b6*1*DIST + v]$
$- [b1 + b2*0 + b3*RMPP + b4*DIST + b5*0*RMPP + b6*0*DIST + v]$
$= \beta 2 + \beta 5 + \beta 6$

We see that the results show a failure to reject the null hypothesis on the clean data in Model 2a and 2b, but comes very close to convincing us on the uncleaned data (model 2) that females do pay higher rents through the mechanisms of RMPP and Distance. This is a clear indicator of the undue influence of the dirty data.

CONCLUSION

The goal of this paper was to show the variety of tools SAS provides to explore your data and to lead to a clean data set on which to test hypothesis about your data using advanced analytics herein exemplified by linear regression. To illustrate how analysis will improve with good EDA, linear regression was run on the dirty and clean data and conclusions were drawn based on the available data. However, this model is limited by the quite small data set and by the amount of missing relevant variables, data that we did not have. If the variables and data set were such that we had more confidence in our results, the next step would be to do robust specification testing of our model. That is not the goal of this paper. Nevertheless, the techniques shown can and should be used in practice on any data set before analysis..

SAS provides many ways of examining our data so we can learn the data generating process, adjust as necessary the manner in which the data is used, create or transform variables to ready the data for analysis. If I made the point that the answer is not as easy as running a hasty regression on unclean data and reporting a quick and default t-test, then I have succeeded and by doing the same you will assure yourself of better and higher quality results.

What is needed are heavy doses of business and economic acumen combined with strong hacking / programming / EDA skills, combined with strong statistical skills. In essence this is what data science is all about.

REFERENCES

Horstman, Joshua. 2018. Doing More with the SGPLOT Procedure, SESUG Paper 205-2018 accessed at https://www.lexjansen.com/sesug/2018/SESUG2018_Paper-205_Final_PDF.pdf.

Kennedy, Peter. 2008. Guide to Econometrics, 6th edition. Wiley-Blackwell.

Myers, Steven C. 2019. Don't let influential data observations kill your regression and your career, manuscript.

Pyndick and Rubinfeld. 1997. Econometric Models & Economic Forecasts, 4th edition. McGraw-Hill

Wicklin, Rich. 2018. Attrs, attrs, everywhere: The interaction between ATTRPRIORITY, CYCLEATTRS, and STYLEATTRS in ODS graphics, The Do Loop blog, accessed at https://blogs.sas.com/content/iml/2018/06/13/attrpriority-cycleattrs-styleattrs-ods-graphics.html.

Wicklin, Rich. 2015. Find the ODS table names produced by any SAS procedure, The Do Loop blog, accessed at https://blogs.sas.com/content/iml/2015/09/08/ods-table-names.html.

SAS 9.4. SAS/STAT User's Guide. The BOXPLOT Procedure. Accessed at https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=statug&docsetTarget=statug_boxplot_details09.htm&locale=en

SAS 9.4. SAS/STAT User's Guide. The LOESS Procedure. Accessed at https://documentation.sas.com/?docsetId=statug&docsetVersion=15.1&docsetTarget=statug_loess_overview.htm&locale=en/

ACKNOWLEDGMENTS

I wish to thank the section chair, Carl Nord, for accepting my paper and for his patience in my completion of it. And my thanks to Jessica Chen, the Academic Chair, for her support as well. I am grateful to Kirk Paul Lafler and Josh Horstman for their encouragement to begin presenting at SAS conferences in general and MWSUG in particular. Finally, I wish to thank the "data scientist" and former student, Brandon Mathias, cited in the introductory section. His story was a perfect way to introduce this paper and helped me keep focus on the big picture of problem solving and not to get lost in the data alone.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Steven C. Myers
myers@uakron.edu
https://www.linkedin.com/in/stevencmyers/
https://econdatascience.com