EEP/IAS 118 - Introductory Applied Econometrics Problem Set 1

Problem Set 1, Spring 2023, Villas-Boas

Due in Gradescope – see deadline due time in Gradescope – Feb 2, 2023

Submit materials (all handwritten/typed answers, Excel notebooks, and R reports) as **one pdf** on Gradescope. For handwritten answers and Excel notebooks, please insert a picture/screenshot directly into this notebook (How to add pictures to the markdown cell (See method 1 or 2)).

After uploading the pdf to Gradescope, please **assign appropriate pages to each question**. Questions that do not have assigned pages on Gradescope may not be graded.

All students currently on the EEP118 bCourses have been added using the bCourses email. If you do not have access to the Gradescope course, please reach out to the GSIs.

Chiara Luna Pilato Moncada / Econometrics / Spring 23 / Problem Set 1

Exercise 1 (Excel / Google Sheets)

Note: Microsoft Office 365 is available from Berkeley Software to students for free - Link.

Relationship between Housing Prices and Violent Crime in 10 US Cities.

We will use September 2021 data from Zumper on one-bedroom apartment prices and 2019 data from the FBI on crime for 10 US cities. The original data has 100 cities. In this first problem set we will only use a subset of the cities. This exercise is to be completed using Excel. We will establish a simple linear relationship between *housing prices and crime* in a subset of cities.

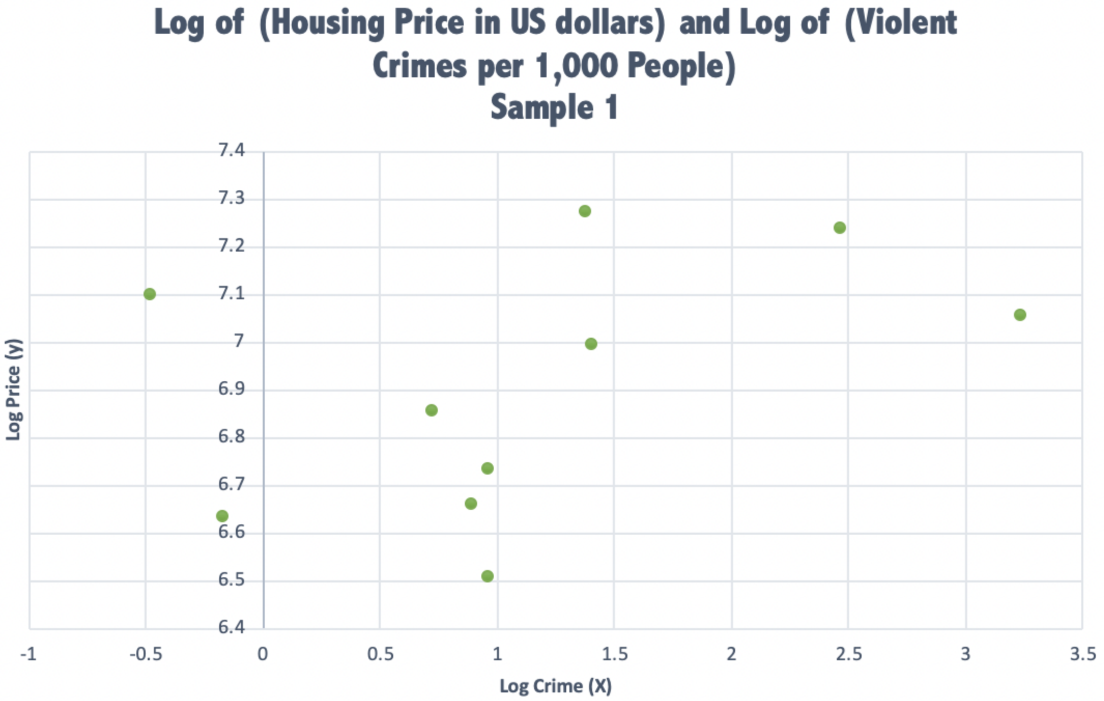
Note: in economics, log always refers to the natural log, ln().

Table 1: Log of (Housing Price in US dollars) and Log of (Violent Crimes per 1,000 People) - Sample 1

CityName (Sample 1)	log of Housing Price (log of Y)	log of Violent Crimes per 1,000 People (log of X)
Arlington	6.85646198	0.72027585
Austin	7.27239839	1.37447478
Corpus Christi	6.73340189	0.96164643
Dallas	7.23705903	2.46504402

CityName (Sample 1)	log of Housing Price (log of Y)	log of Violent Crimes per 1,000 People (log of X)
El Paso	6.65929392	0.88459365
Fort Worth	6.99393298	1.40315148
Houston	7.05617528	3.22910335
Irving	7.09837564	-0.4828863
Laredo	6.63331843	-0.1791267
Lubbock	6.50727771	0.96049899

(a) Use Excel to create a scatter plot of these observations. Don't forget to (1) label the axes and their units, and (2) title your graph. **You should use the tables provided here for these calculations, not the actual observations from the .csv data file.**

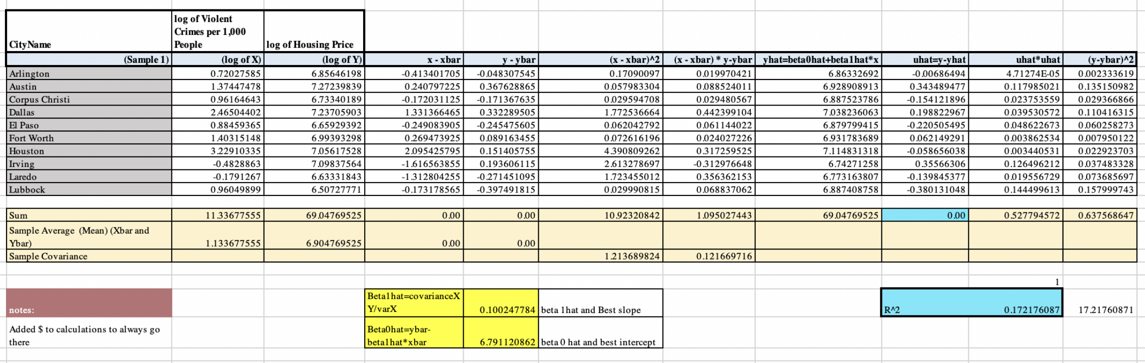


(b) This question has **two parts**.

First: Estimate the linear relationship between the log of Housing Price (log(Y)) and the log of violent crimes per 1,000 people (log(X)) by OLS, showing all intermediate calculations as we saw in the lecture 3 slides (use Excel to create the table and show all the steps).

Second: interpret the value of the estimated parameters β_0 and β_1.

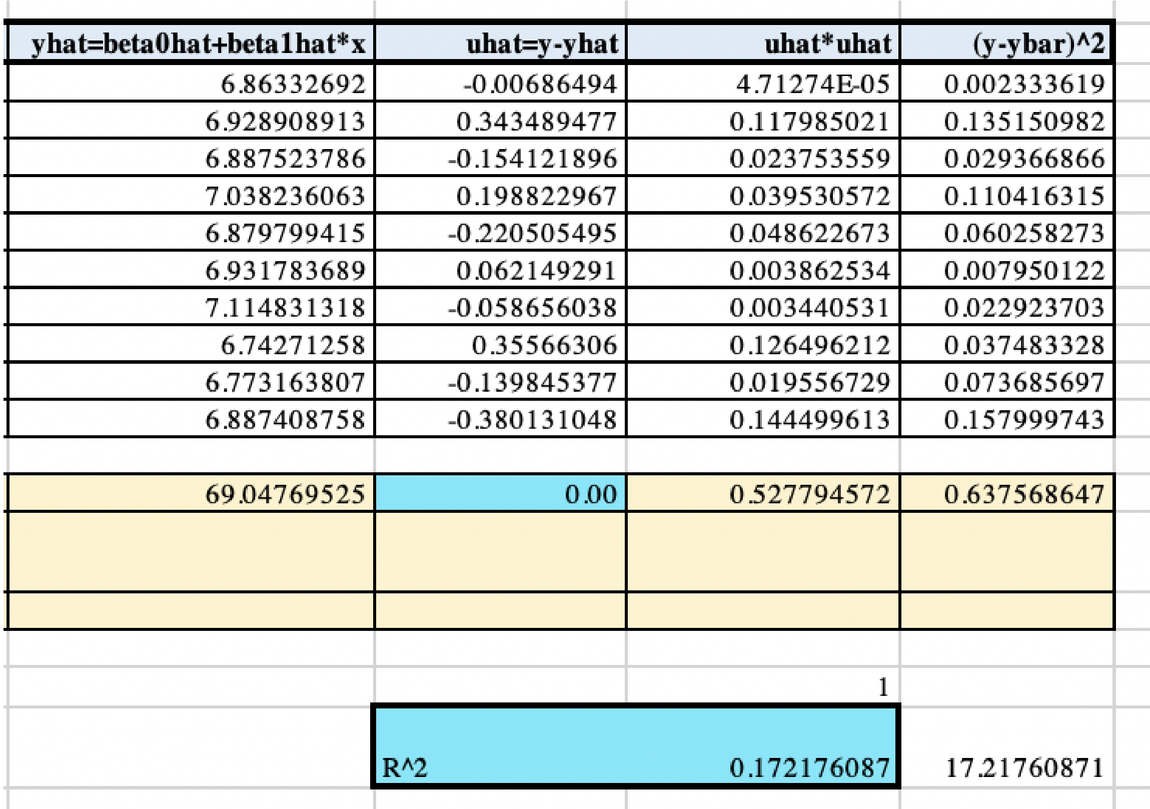
$$\widehat{log(Y_i)} = \hat{\beta}_0 + \hat{\beta}_1 log(X_i) \qquad \text{i} = \{\text{cities in sample 1}\}$$



CityName	log of Violent Crimes per 1,000 People	log of Housing Price								
(Sample 1)	(log of X)	(log of Y)	x - xbar	y - ybar	(x - xbar)^2	(x - xbar) * y-ybar	yhat=beta0hat+beta1hat*x	uhat=y-yhat	uhat*uhat	(y-ybar)^2
Arlington	0.72027585	6.85646198	-0.413401705	-0.048307545	0.17090097	0.019970421	6.86332692	-0.00686494	4.71274E-05	0.002333619
Austin	1.37447478	7.27239839	0.240797225	0.367628865	0.057983304	0.088524011	6.928908913	0.343489477	0.117985021	0.135150982
Corpus Christi	0.96164643	6.73340189	-0.172031125	-0.171367635	0.029594708	0.029480567	6.887523786	-0.154121896	0.023753559	0.029366866
Dallas	2.46504402	7.23705903	1.331366465	0.332289505	1.772536664	0.442399104	7.038236063	0.198822967	0.039530572	0.110416315
El Paso	0.88459365	6.65929392	-0.249083905	-0.245475605	0.062042792	0.061144022	6.879799415	-0.220505495	0.048622673	0.060258273
Fort Worth	1.40315148	6.99393298	0.269473925	0.089163455	0.072616196	0.024027226	6.931783689	0.062149291	0.003862534	0.007950122
Houston	3.22910335	7.05617528	2.095425795	0.151405755	4.390809262	0.317259525	7.114831318	-0.058656038	0.003440531	0.022923703
Irving	-0.4828863	7.09837564	-1.616563855	0.193606115	2.613278697	-0.312976648	6.74271258	0.35566306	0.126496212	0.037483328
Laredo	-0.1791267	6.63331843	-1.312804255	-0.271451095	1.723455012	0.356362153	6.773163807	-0.139845377	0.019556729	0.073685697
Lubbock	0.96049899	6.50727771	-0.173178565	-0.397491815	0.029990815	0.068837062	6.887408758	-0.380131048	0.144499613	0.157999743
Sum	11.33677555	69.04769525	0.00	0.00	10.92320842	1.095027443	69.04769525	0.00	0.527794572	0.637568647
Sample Average (Mean) (Xbar and Ybar)	1.133677555	6.904769525	0.00	0.00						
Sample Covariance					1.213689824	0.121669716				
									1	
notes:			Beta1hat=covarianceX Y/varX	0.100247784	beta 1hat and Best slope			R^2	0.172176087	17.21760871
Added $ to calculations to always go there			Beta0hat=ybar-beta1hat*xbar	6.791120862	beta 0 hat and best intercept					

Second part: After doing the calculations for Beta0hat and Beta1hat I got the following best values. For Beta1hat, the value is approximately 0.10024, as the slope of the regression line of our scatter plot. I also got that Beta0hat has the approximate value of 6.7911 as the Y intercept for the regression line of the scatter plot

(c) In your table, compute the fitted value and the residual for each observation, and verify that the residuals (approximately) sum to 0.



yhat=beta0hat+beta1hat*x	uhat=y-yhat	uhat*uhat	(y-ybar)^2
6.86332692	-0.00686494	4.71274E-05	0.002333619
6.928908913	0.343489477	0.117985021	0.135150982
6.887523786	-0.154121896	0.023753559	0.029366866
7.038236063	0.198822967	0.039530572	0.110416315
6.879799415	-0.220505495	0.048622673	0.060258273
6.931783689	0.062149291	0.003862534	0.007950122
7.114831318	-0.058656038	0.003440531	0.022923703
6.74271258	0.35566306	0.126496212	0.037483328
6.773163807	-0.139845377	0.019556729	0.073685697
6.887408758	-0.380131048	0.144499613	0.157999743
69.04769525	0.00	0.527794572	0.637568647
		1	
	R^2	0.172176087	17.21760871

In my excell, I did get that the sum of the residuals summed up to 0, I had to fix the way the value was shown although, becasue Excell was not rounding up the number and I got a very large E value. Once I addressed that code, the table did show me the value of 0

(d) According to the estimated relation, what is the predicted $\hat{Y}$ (**level**, not log of prices) for a city with a log crime of -5.2? (Pay attention to units)

Here its important to note we are using a Log-Log equation, so according to the formula we use for logyhat = beta0hat+beta1hat(x), we get the following answer, approximately: logyhat = beta0hat+beta1hat(-5.2) = 6.26983. Doing e^6.26983 to solve for logyhat then would give us 528.38 as the predicted y hat

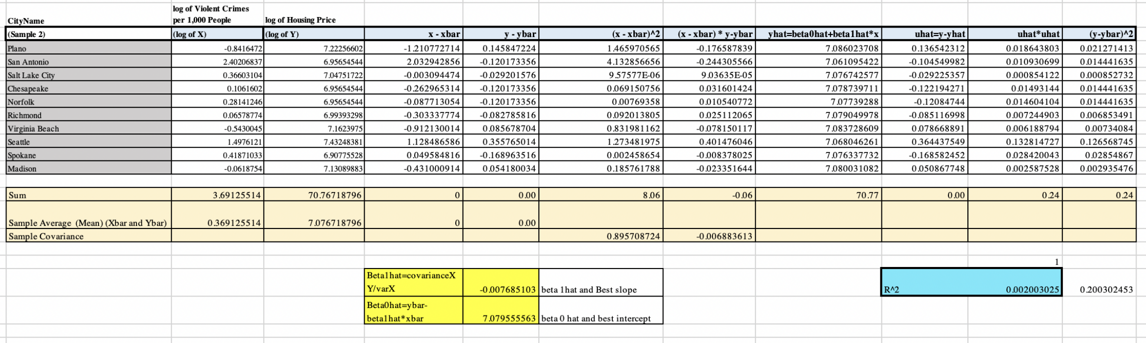
(e) How much of the variation in per capita log home prices in these 10 cities is explained by the log of violent crimes per 1,000 people?

As seen in the excell sheet, the R^2, which is how much of the variation we can "explain away" in the model, is approximetly 0.1721 or 17.217%

(f) Repeat exercise (b) for one additional set of 10 cities below. **You should use Table 2 provided above for these calculations, not the actual observations from the .csv data file.**

Table 2: Log of (Housing Price in US dollars) and Log of (Violent Crimes per 1,000 People) - Sample 2

CityName (Sample 2)	log of Housing Price (log of Y)	log of Violent Crimes per 1,000 People (log of X)
Plano	7.22256602	-0.8416472
San Antonio	6.95654544	2.40206837
Salt Lake City	7.04751722	0.36603104
Chesapeake	6.95654544	0.1061602
Norfolk	6.95654544	0.28141246
Richmond	6.99393298	0.06578774
Virginia Beach	7.1623975	-0.5430045
Seattle	7.43248381	1.4976121
Spokane	6.90775528	0.41871033
Madison	7.13089883	-0.0618754



CityName	log of Violent Crimes per 1,000 People	log of Housing Price									
(Sample 2)	(log of X)	(log of Y)	x - xbar	y - ybar	(x - xbar)^2	(x - xbar) * y-ybar	yhat=beta0hat+beta1hat*x	uhat=y-yhat	uhat*uhat	(y-ybar)^2	
Plano	-0.8416472	7.22256602	-1.210772714	0.145847224	1.465970565	-0.176587839	7.086023708	0.136542312	0.018643803	0.021271413	
San Antonio	2.40206837	6.95654544	2.032942856	-0.120173356	4.132856656	-0.244305566	7.061095422	-0.104549982	0.010930699	0.014441635	
Salt Lake City	0.36603104	7.04751722	-0.003094474	-0.029201576	9.57577E-06	9.03635E-05	7.076742577	-0.029225357	0.000854122	0.00852732	
Chesapeake	0.1061602	6.95654544	-0.262965314	-0.120173356	0.069150756	0.031601424	7.078739711	-0.122194271	0.01493144	0.014441635	
Norfolk	0.28141246	6.95654544	-0.087713054	-0.120173356	0.00769358	0.010540772	7.07739288	-0.12084744	0.014604104	0.014441635	
Richmond	0.06578774	6.99393298	-0.303337774	-0.082785816	0.092013805	0.025112065	7.079049978	-0.085116998	0.007244903	0.006853491	
Virginia Beach	-0.5430045	7.1623975	-0.912130014	0.085678704	0.831981162	-0.078150117	7.083728609	0.078668891	0.006188794	0.00734084	
Seattle	1.4976121	7.43248381	1.128486586	0.355765014	1.273481975	0.401476046	7.068046261	0.364437549	0.132814727	0.126568745	
Spokane	0.41871033	6.90775528	0.049584816	-0.168963516	0.002458654	-0.008378025	7.076337732	-0.168582452	0.028420043	0.02854867	
Madison	-0.0618754	7.13089883	-0.431000914	0.054180034	0.185761788	-0.023351644	7.080031082	0.050867748	0.002587528	0.002935476	
Sum	3.69125514	70.76718796	0	0.00	8.06	-0.06	70.77	0.00	0.24	0.24	
Sample Average (Mean) (Xbar and Ybar)	0.369125514	7.076718796	0	0.00							
Sample Covariance					0.895708724	-0.006883613					
									1		
			Beta1hat=covarianceX Y/varX	-0.007685103	beta 1hat and Best slope			R^2	0.002003025	0.200302453	
			Beta0hat=ybar-beta1hat*xbar	7.079555563	beta 0 hat and best intercept						

Second part: After doing the calculations for Beta0hat and Beta1hat I got the following best values. For Beta1hat, the value is approximately -0.007685, as the slope of the regression line of our scatter plot. I also got that Beta0hat has the approximate value of 7.07955 as the Y intercept for the regression line of the scatter plot

(g) Do your estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ change between Tables 1, and 2? Why?

Yes, my estimates of Beta0hat and Beta1hat vary between table 1 and table 2. This could be because in these samples we have small sizes of only 10. Because these samples are small, there will always be space for variation between the answers. The bigger the sample sizes we use in experiments, the higher accuracy we will have.

Exercise 2 (Functional Forms)

Suppose you estimate alternative specifications as given below *using data from 41 cities*:

A linear relationship: $\hat{Y}_i = 996 + 29.6X_i$

A linear-log relationship: $\hat{Y}_i = 1065 + 90.8\log(X_i)$

A log-log relationship: $\widehat{\log(Y_i)} = 6.97 + 0.05\log(X_i)$

Note that it is convention to always use the natural log.

(i) Interpret the parameter on violent crimes X (or log of violent crimes per 1000 people $\log(X)$) in each of these equations.

For the **Linear Relationship**, we see that the intercept in this case would be 996 and the slope would be 29.6 Interpreting this we get that a 1 unit increase in violent crimes X would lead (is expected to lead) to a 29.6 unit increase in y (Housing Prices).

For the **Linear Log Relationship**, we see the slope is 90.8 and the intercept would be 1065. Interpreting this we get that a 1% increase in violent crimes x would lead (is expected to lead) to a (90.8/100) unit increase in y. Which would be a 0.908 unit increase in y (Housing Prices). In this case, we would not multiply the result by any other percent change in x, as we are not given any specific % change.

For the **log-log relationship**, we see the slope is 0.05log and the interncept in this case would be 6.97. Interpreting this we get that a 1% increase in violent crimes x is expected to lead to a 0.05% increase in y(Housing Prices). Just like the previous one, because we dont have any other additional information, we dont do any other procedures on this solution.

(ii) What is the predicted one bedroom rental price in dollars for a city with a crime per 1000 people equal to 5.2 in each of these equations?

For the Linear relationship, we have Yi^= 996+29.6(5.2) = 1,149.92

For the Linear Log Relationship, we have Yi^=1065+90.8log(5.2) = 1130.01

For the Log Log relationship, we have log(Yi)ˆ=6.97+0.05log(5.2). This gave me log(Yi)ˆ=7.0058, which becomes e^7.0058 = 1,103.01

Exercise 3. Importing data into R and Basic R first commands

For the purposes of this class, we will be using RStudio or a cloud-based version of RStudio provided through UC Berkeley's $Datahub$. The data files can be accessed directly through $Datahub$ and do not require you to install anything on your computer. This exercise is designed to get you familiar with R (if you have RStudio on your computer) or accessing the service.

The exercise will have you learn about loading data and obtaining summary statistics. To start off, we're going to use Jupyter notebooks to help familiarize you with some R commands. For help with Jupyter and R, refer to the Coding Bootcamp Part 1 recording on bCourses and the corresponding interactive notebook on $Datahub$.

*Note: Coding Bootcamp Part 1 covers all necessary R methods.

(a) To access the Jupyter notebook for this problem set on Datahub, click the following link and navigate to the Problem Set 1 folder.

Skip! You are already here - nice work.

(b) Load the datafile "dataPset1_2023.csv" into R (since this is a ".csv" file, you should use the read.csv() function).

In [1]:
```
ps1_df<- read.csv("dataPset1_2023.csv")
head(ps1_df)
```

```
                     A data.frame: 6 × 6
        city  state  pricesept2021  violentcrime  logPrice   logCrime
       <chr>  <chr>          <int>         <dbl>     <dbl>      <dbl>
1  Henderson     NV           1490         0.543  7.306531 -0.6106460
2  Las Vegas     NV           1170         8.854  7.064759  2.1808693
3       Reno     NV           1250         1.419  7.130899  0.3499524
4    Buffalo     NY           1050         2.533  6.956545  0.9294044
5   New York     NY           2950        47.821  7.989560  3.8674649
6  Rochester     NY            980         1.540  6.887553  0.4317824
```

(c) Provide basic summary statistics on the log of home price (logPrice) in the dataframe. Use the summary() command. This function is part of base R, so you do not need to load any packages before using it. What is the median value of log housing price in cities in the sample?

In [5]:
```
summary(ps1_df)
```

```
     city               state           pricesept2021   violentcrime
 Length:42          Length:42          Min.   : 640   Min.   : 0.120
 Class :character   Class :character   1st Qu.: 925   1st Qu.: 1.161
 Mode  :character   Mode  :character   Median :1090   Median : 2.478
                                       Mean   :1149   Mean   : 5.173
                                       3rd Qu.:1312   3rd Qu.: 4.704
                                       Max.   :2950   Max.   :47.821
    logPrice        logCrime
 Min.   :6.461   Min.   :-2.1203
 1st Qu.:6.830   1st Qu.: 0.1486
 Median :6.994   Median : 0.9070
 Mean   :7.005   Mean   : 0.9337
 3rd Qu.:7.180   3rd Qu.: 1.5482
 Max.   :7.990   Max.   : 3.8675
```

In [2]:

```
summary(ps1_df$logPrice)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.461   6.830   6.994   7.005   7.180   7.990 
```

Here, I summarized both the general data, and then also just the individual data that pertains to logPrice. In this case, as seen in the data, the median log housing price in the sample is 6.994

(d) Next, generate custom summary statistics on the Log of violent crime (logCrime) using the summarise() function provided by dplyr. At minimum, your summary stats should include (a) obs. count, (b) min, mean, median, max, (c) range, and (d) std. deviation. You will need to call the tidyverse package with the library() function to use it (tidyverse is a collection of packages designed for data science. It includes dplyr and several other packages we'll use this term).

In [5]:

```
library(tidyverse)
library(haven)

ps1_df%>%
summarise(obs_count = n(),
          min=min(logCrime),
          mean=mean(logCrime),
          median=median(logCrime),
          max=max(logCrime),
          range=range(logCrime),
          sd=sd(logCrime))
```

A data.frame: 2 × 7

obs_count	min	mean	median	max	range	sd
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
42	-2.120264	0.9337299	0.906999	3.867465	-2.120264	1.180533
42	-2.120264	0.9337299	0.906999	3.867465	3.867465	1.180533

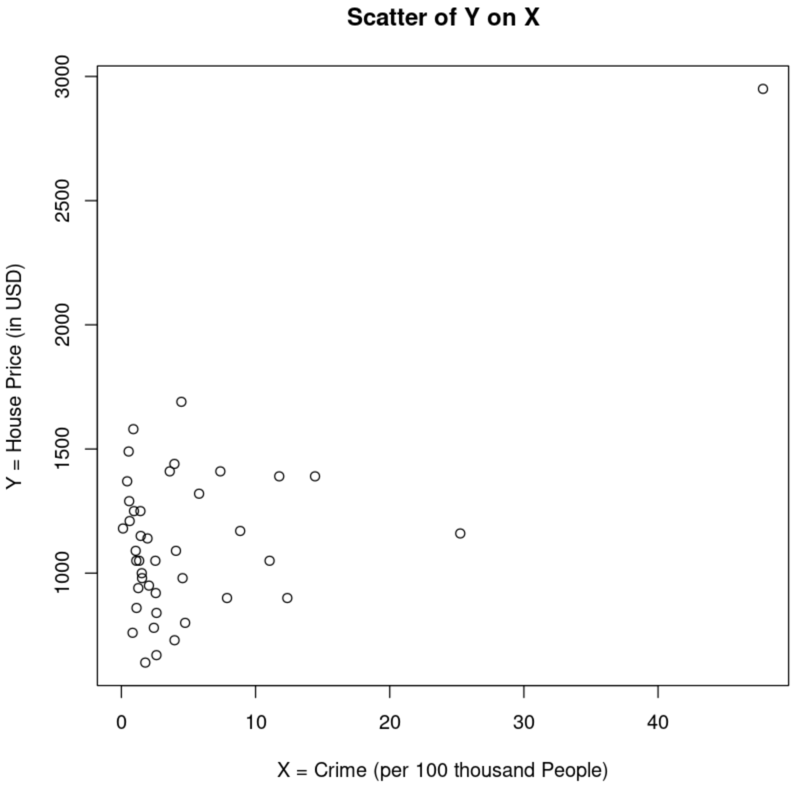
Here we see that the obs count is 42. The Min is -2.12, the mean is 0.933, the median is 0.90699, the max is 3.8674, the range is from -2.1202 to 3.8674, and the standard deviation is 1.180

(e) Create a scatter plot of the Price and Crime data in levels. Use

```
fig1 <- plot(ps1_df$violentcrime, ps1_df$pricesept2021, main="Scatter of Y on X", xlab="X = Crime (per 100 thousand People)", ylab="Y = House Price (in USD)")
```

Note: Make sure to print the scatterplot in the notebook's codecell.

```
In [7]: fig1 <- plot(ps1_df$violentcrime, ps1_df$pricesept2021,
                     main="Scatter of Y on X",
                     xlab="X = Crime (per 100 thousand People)",
                     ylab="Y = House Price (in USD)")
```



(f) Save a pdf to your computer. This can be done by going to `File > Save and Export Notebook As... > PDF`.

Note: Make sure to check your pdf before uploading to ensure all the code cells are run and all text/output is visible, and insert all screenshots from Excel/handwritten work before uploading to Gradescope.