Tutorial 1

Statistics 2 for IBA

Tilburg University

Introduction

- In this tutorial we will explore the `f150.sav` dataset, which featured in the Resit exam in January 2021.
- The dataset contains information from advertisements for second-hand Ford F-150 trucks on Craigslist, a US website similar to Marktplaats here in NL.
- The data contains the asking price, the year of manufacture, the mileage (odometer reading), the truck's color and an indicator for whether truck is in good condition or not.

Dataset Description as seen in TestVision

Introduction

The data set for this exam can be downloaded by clicking on the following link: f150

Craigslist is a popular website in the United States where people can place classified advertisements to sell second-hand items. It is similar to *marktplaats.nl* in the Netherlands. Many second-hand cars are advertised on the website. The most common vehicle advertised on the website is the Ford F-150 pickup truck. Here is an example of what a Ford F-150 pickup truck looks like:

You are interested in understanding how characteristics of a second-hand truck, such as its age and mileage, affect the price it can sell for. You have data on 500 different Ford F-150 advertisements on Craigslist, where for each advertisement you observe the following variables:

price	The asking price of the truck
year	The year the truck was purchased new
odometer	The total number of miles the truck has driven
paint_color	The color of the truck (character/string variable)
good_condition	Dummy variable for if the truck is in good condition or not (=1 for good condition and =0 for not in good condition).

The advertisements were collected at the end of the year in 2020. Therefore if *year* is equal to 2019, the *age* of the truck is 1 year. If *year* is equal to 2018, the *age* of the truck is 2 years, and so on.

Exercises

1. Open the `f150.sav` data file and inspect the data. Look at *Data View* and *Variable View*.
2. Create a histogram of the variable *Price* using `Graphs`→`Legacy Dialogs`→`Histogram`.
3. Obtain descriptive statistics (mean/min/max/SD) of the variable *Price*.
4. Create a scatter plot of *Year* against *Price*, with *Year* on the horizontal axis and price on the vertical axis. Interpret it.
5. Compute the covariance and correlation between *Year* and *Price*. Interpret them.
6. Create the variable *Age* from *Year*. Keep in mind that the advertisements were shown in 2020.
7. Compute the covariance and correlation between *Age* and *Price*. Relate this to what you found in Q5.
8. Study the relationship between *Age* and *Year*. Create a scatter plot and compute the covariance and correlation. Interpret them.

Bonus Questions: Q1 from the January 2021 Resit

What is the sample correlation coefficient between *price* and *odometer*?

123

You were asked to type a number into the box.

Bonus Questions: Q2 from the January 2021 Resit

Choose the answer below which best describes the interpretation of the sample correlation coefficient between *price* and ***year.***

Note: See the block intro for how the *year* of an F-150 relates to its *age*.

- ○ There is a negative linear relationship between the *age* of an F-150 and its price.
- ○ A one-year increase in the *age* of an F-150 on average increases its price.
- ○ There is no relationship between the price of an F-150 and its *age*.
- ○ There is a positive linear relationship between the *age* of an F-150 and its price.

Multiple choice question.

Q1: Data View

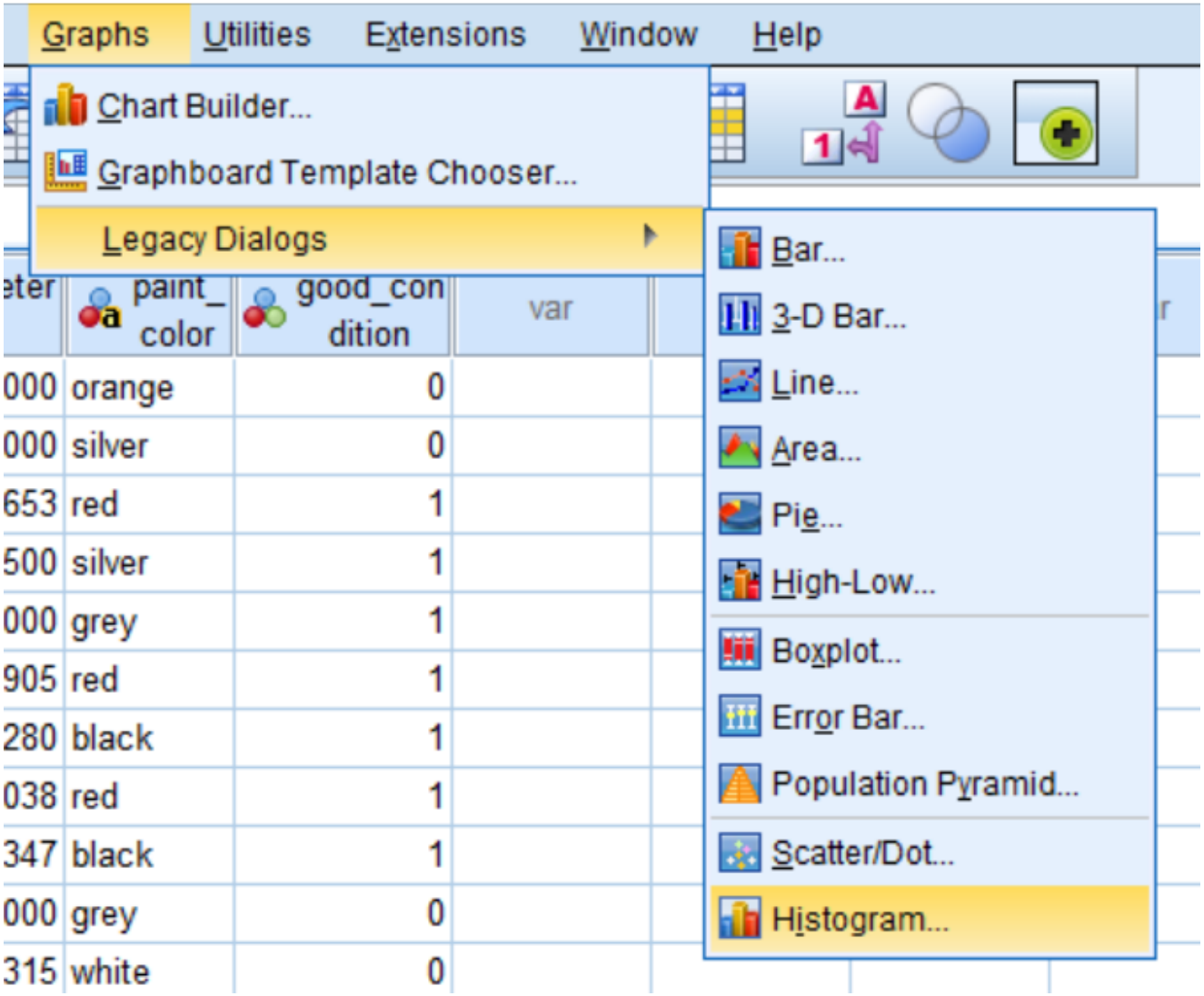
	price	year	odometer	paint_color	good_condition
1	3000	1979	150000	orange	0
2	3200	1993	190000	silver	0
3	37800	2017	52653	red	1
4	16450	2016	49500	silver	1
5	7500	2008	240000	grey	1
6	19900	2016	105905	red	1
7	31995	2016	75280	black	1
8	11140	2006	114038	red	1
9	18500	2013	119347	black	1
10	3900	2003	160000	grey	0

Q1: Variable View

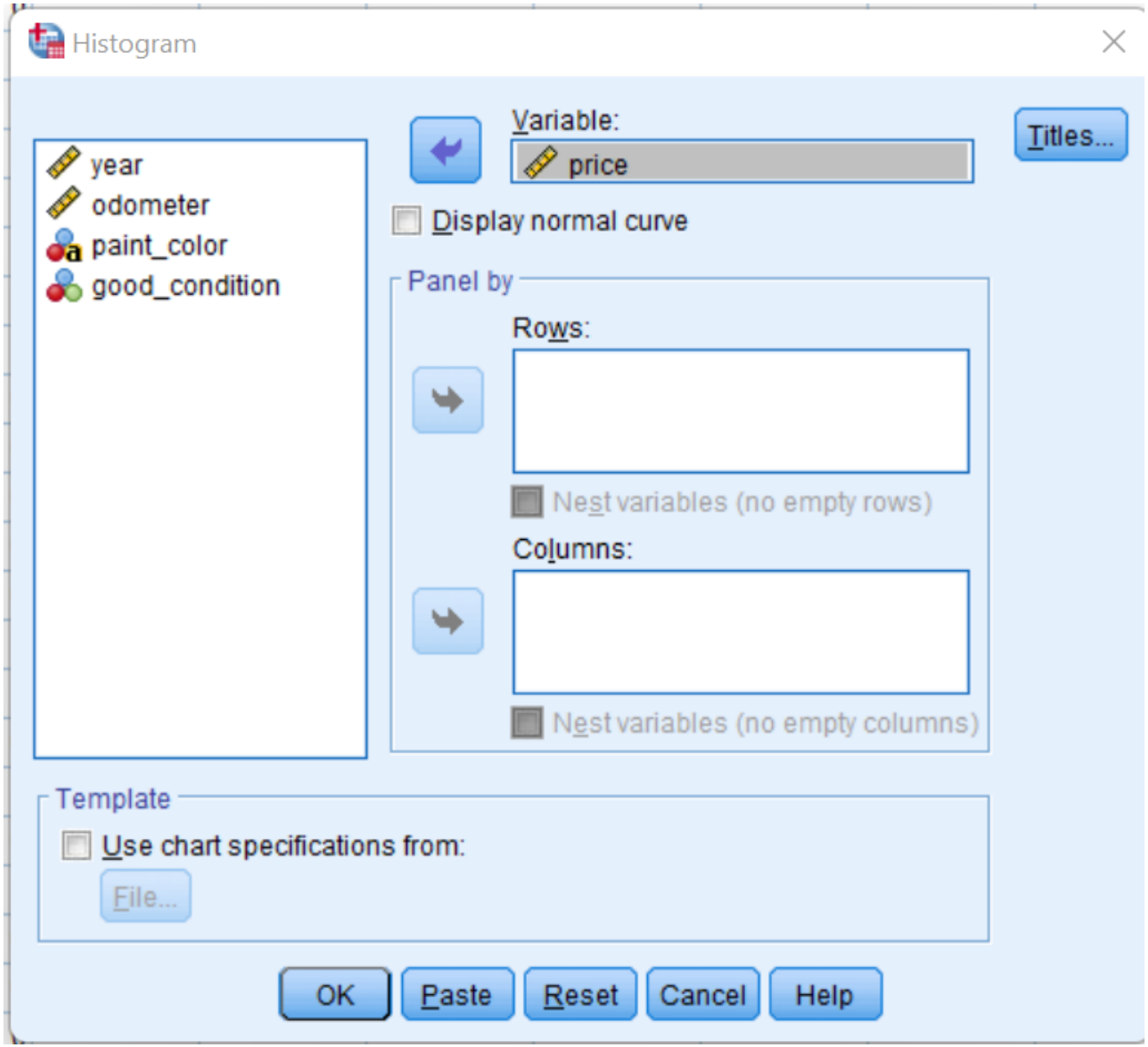
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	price	Numeric	5	0		None	None	8	Right	Scale	Input
2	year	Numeric	4	0		None	None	8	Right	Scale	Input
3	odometer	Numeric	6	0		None	None	8	Right	Scale	Input
4	paint_color	String	6	0		None	None	6	Left	Nominal	Input
5	good_condit...	Numeric	1	0		None	None	8	Right	Nominal	Input

- Price, year and odometer are continuous numerical variables.
- Paint color is a *string* (character) variables.
- Good condition is an indicator (dummy) variable: it equals 1 when the truck is in good condition and 0 when it is not.

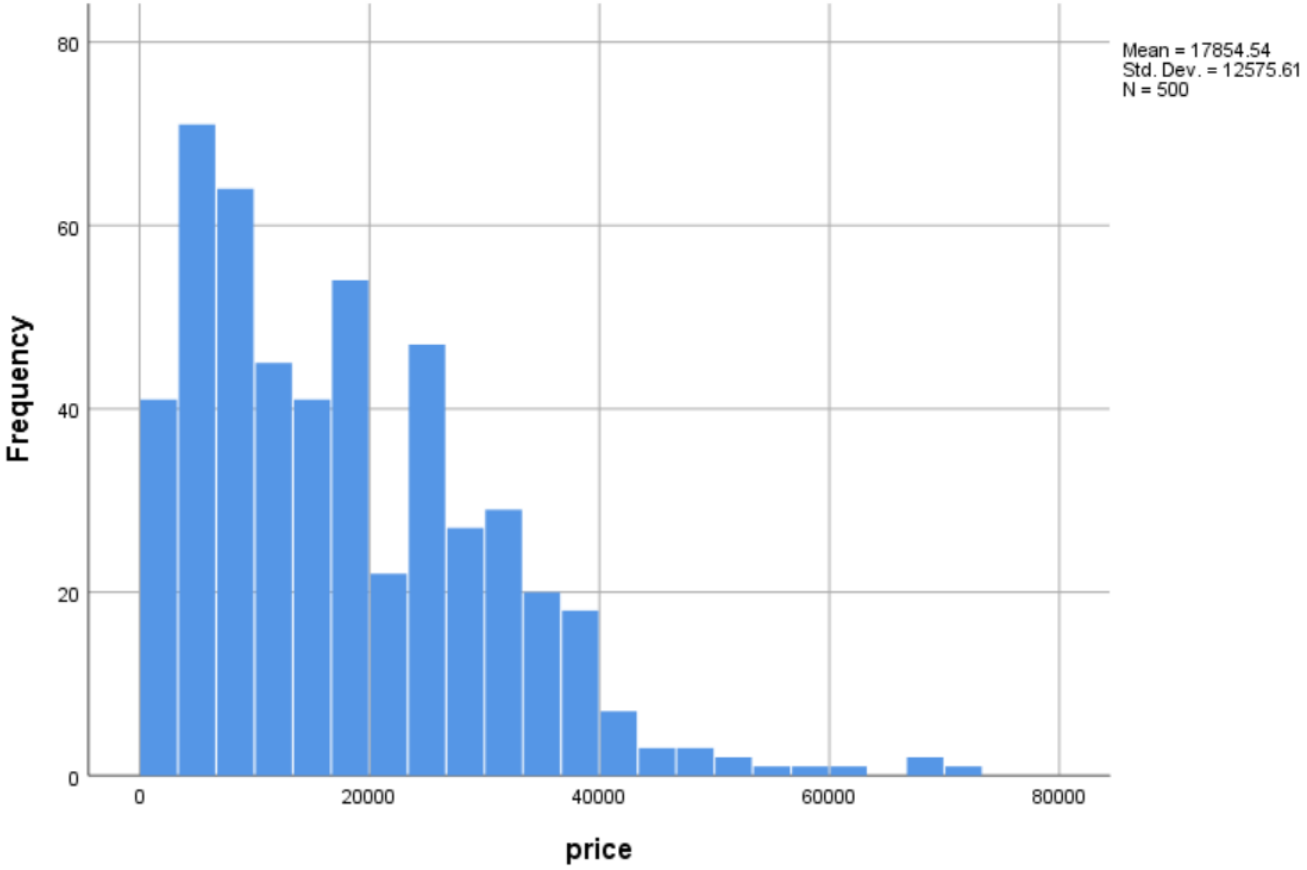
Q2: Creating a Histogram



Q2: Creating a Histogram

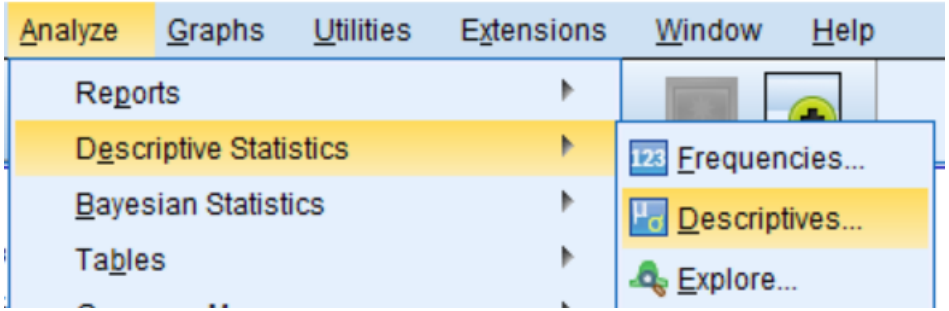


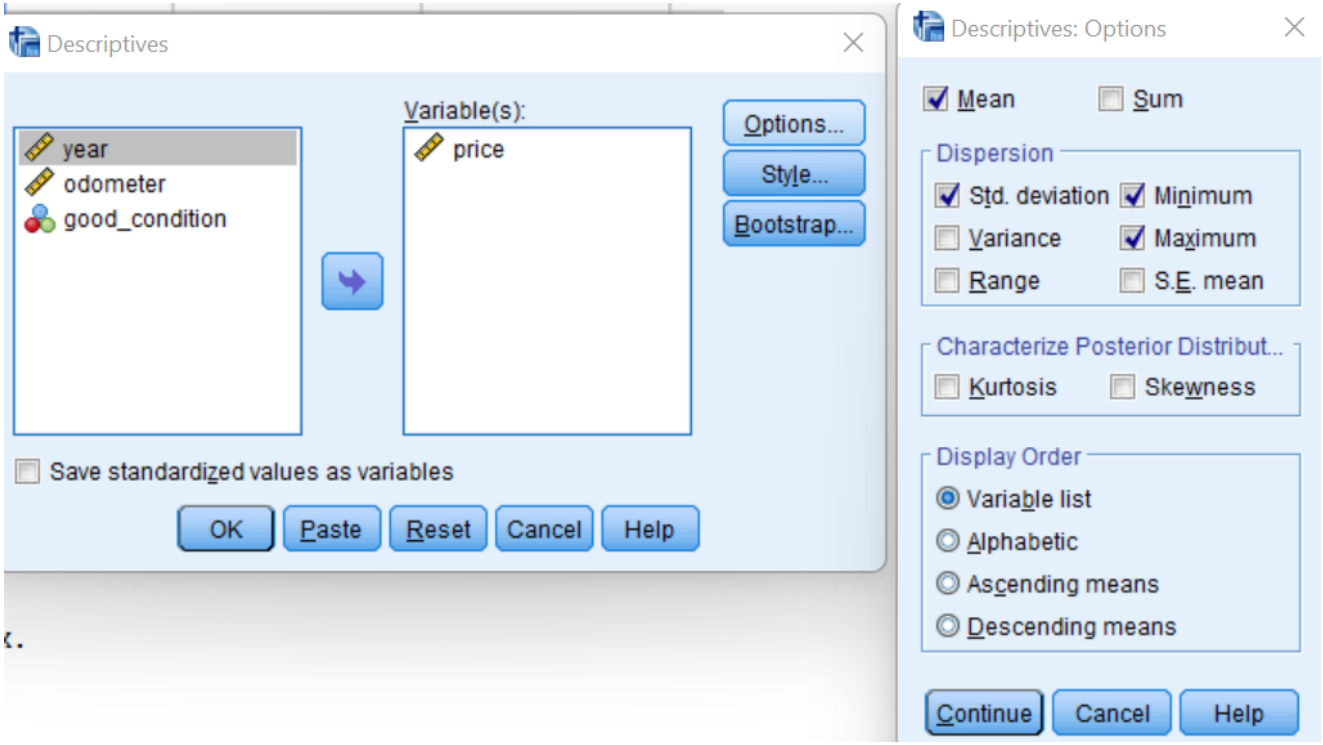
Q2: Creating a Histogram



- Prices vary between (slightly above) zero and about 70k.
- Most prices are below 40k.
- The distribution is skewed to the right.

Q3: Obtaining Descriptive Statistics





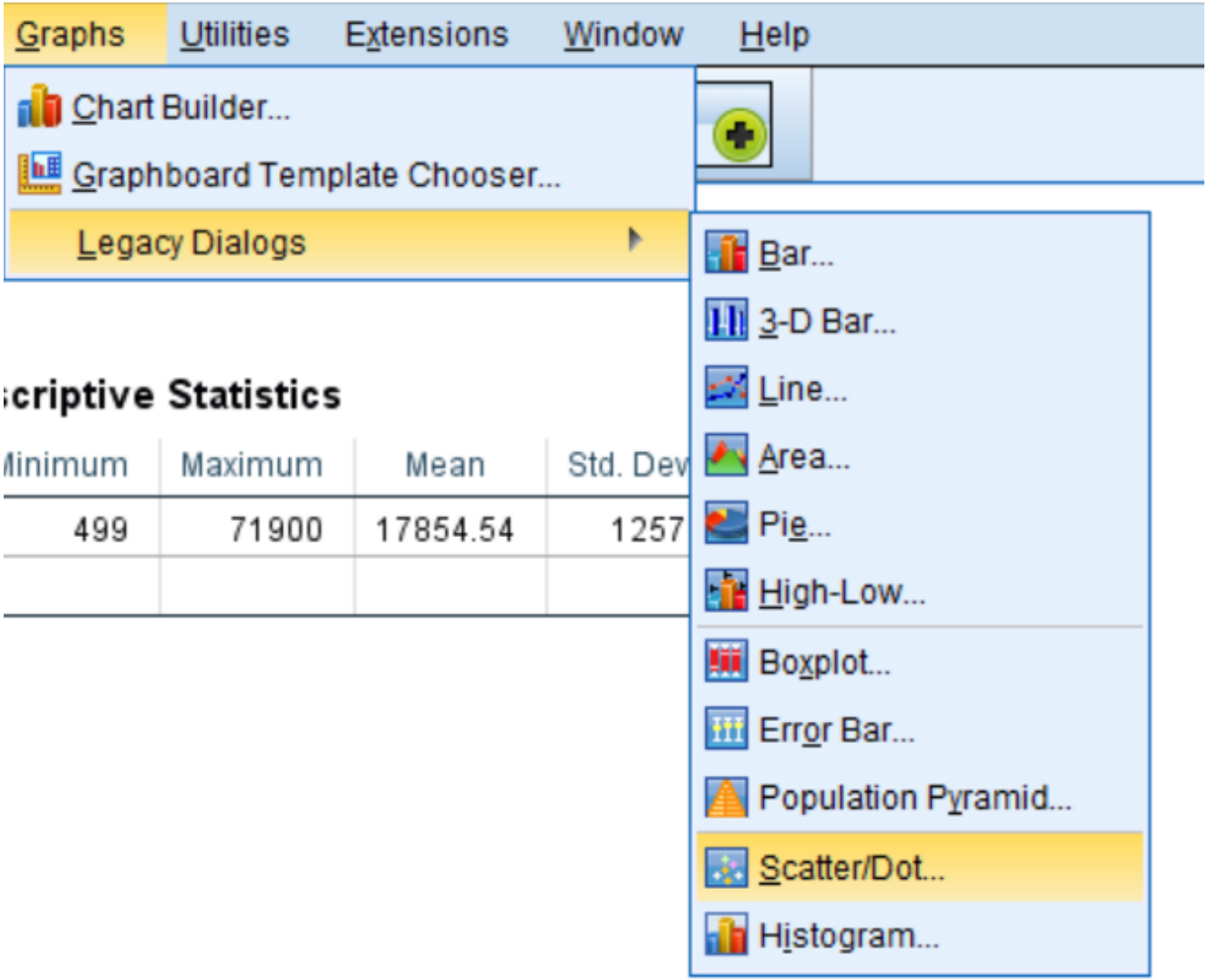
Q3: Obtaining Descriptive Statistics

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
price	500	499	71900	17854.54	12575.610
Valid N (listwise)	500				

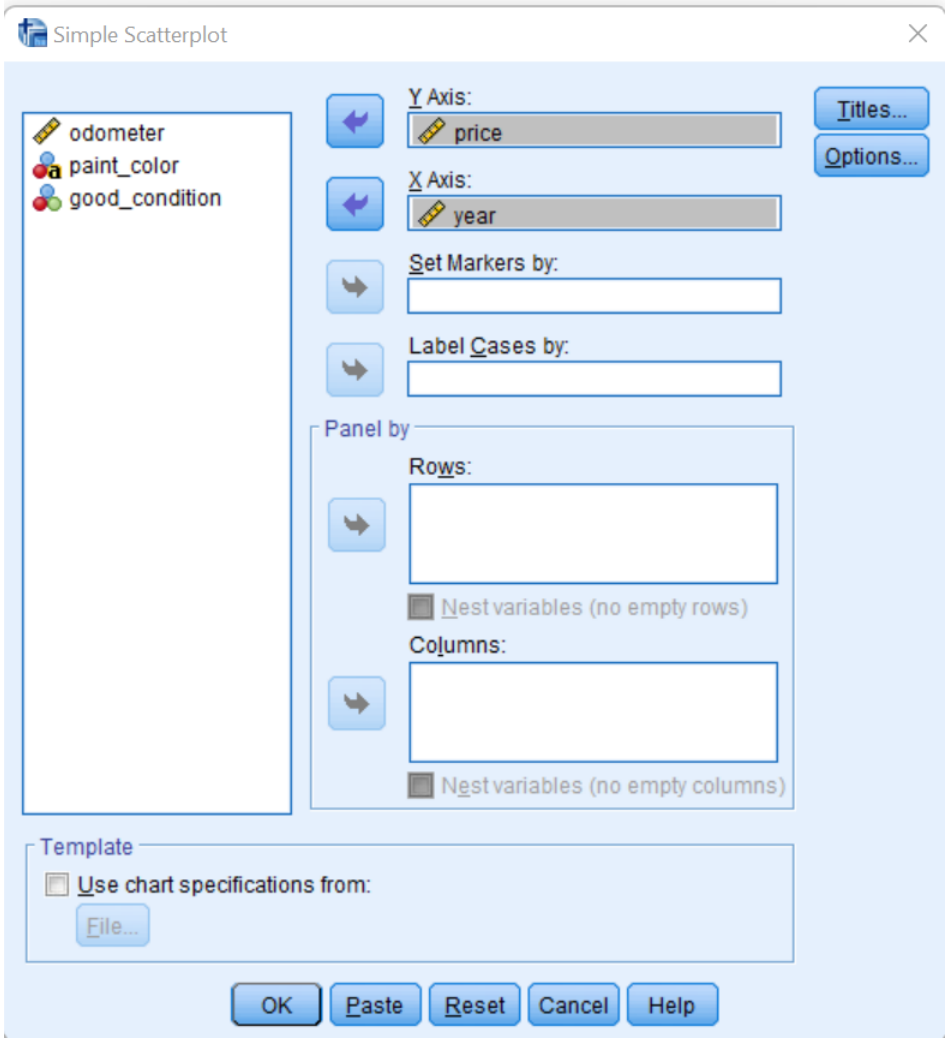
- 500 observations in total.
- The lowest advertised price is \$499, the highest is \$71,900.
- The average is \$17,854.54.
- The standard deviation is \$12,575.61.

Q4: Creating a Scatter Plot

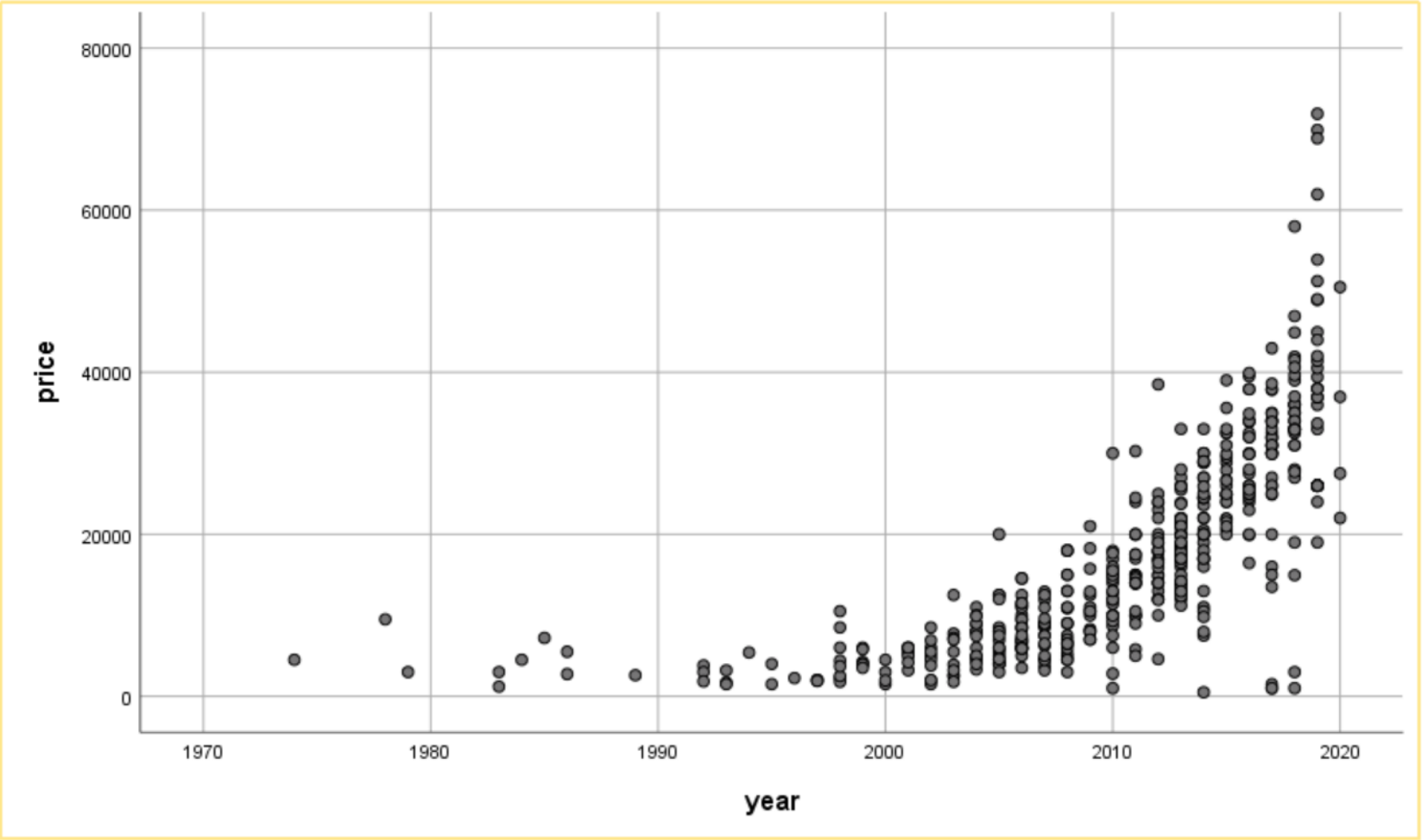


Q4: Creating a Scatter Plot

Put variable on the horizontal axis in the x-axis box.
Put variable on the vertical axis in the y-axis box.

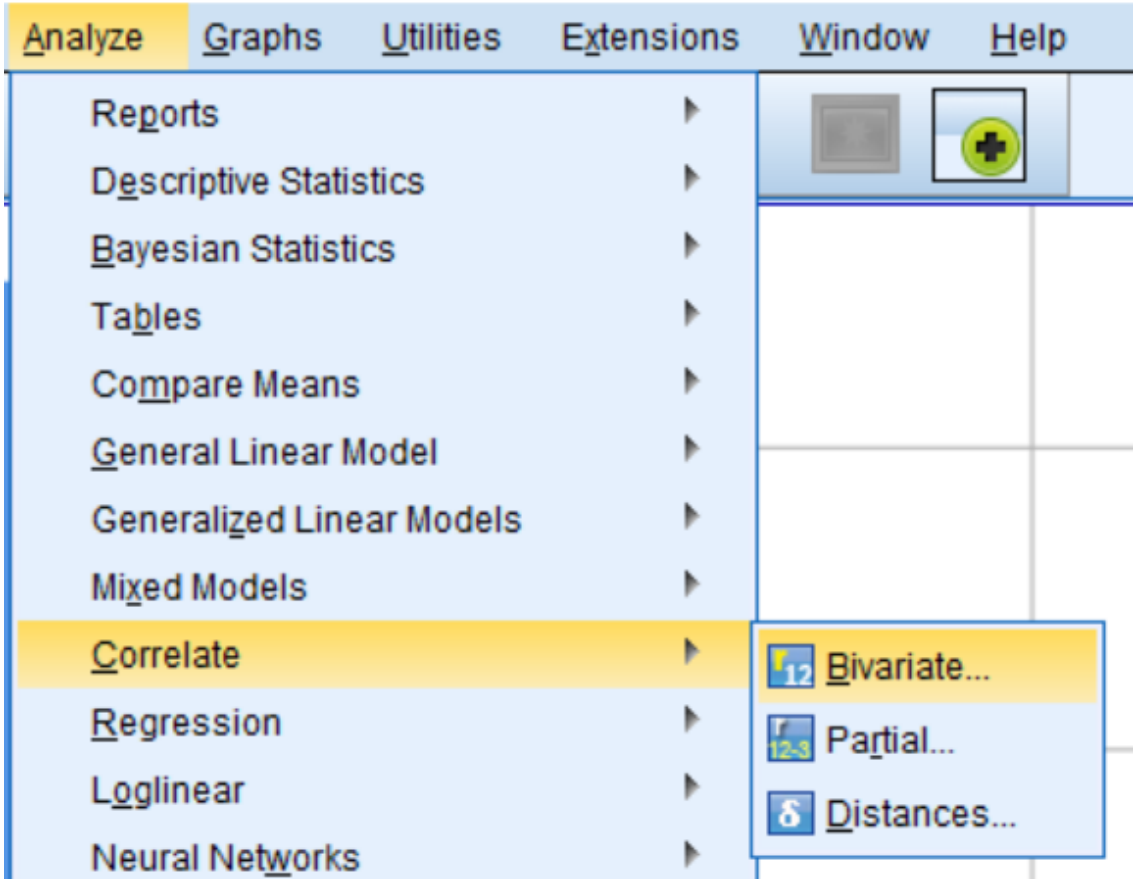


Q4: Creating a Scatter Plot



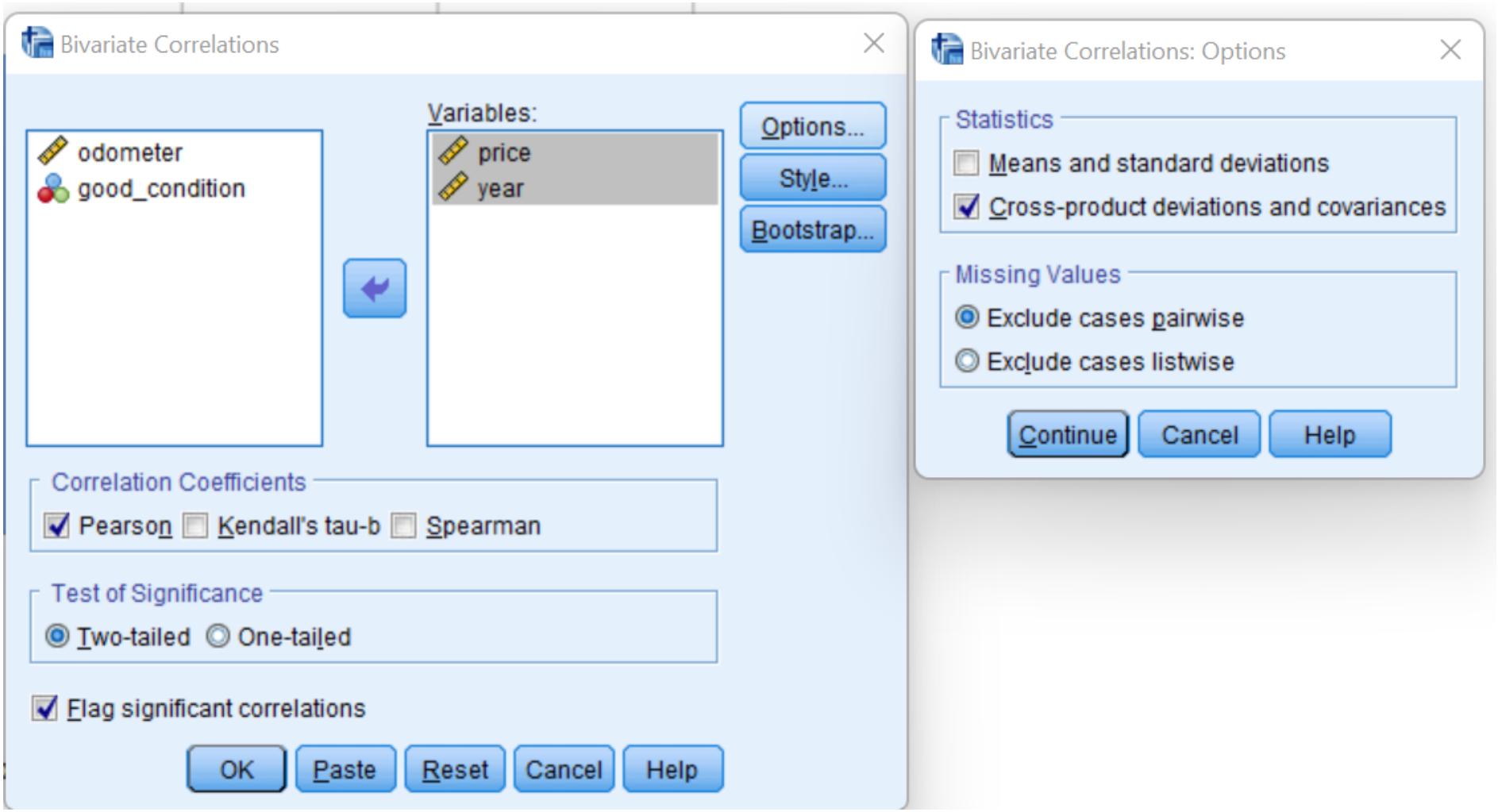
- Positive relationship between year and price.
- Somewhat nonlinear relationship: cars from before 2000 have a low price, but starting 2000 onwards the relationship is linear.

Q5: Covariance and Correlation



Q5: Covariance and Correlation

Click `Options...` to add covariances to the table.



Q5: Covariance and Correlation

➜ **Correlations**

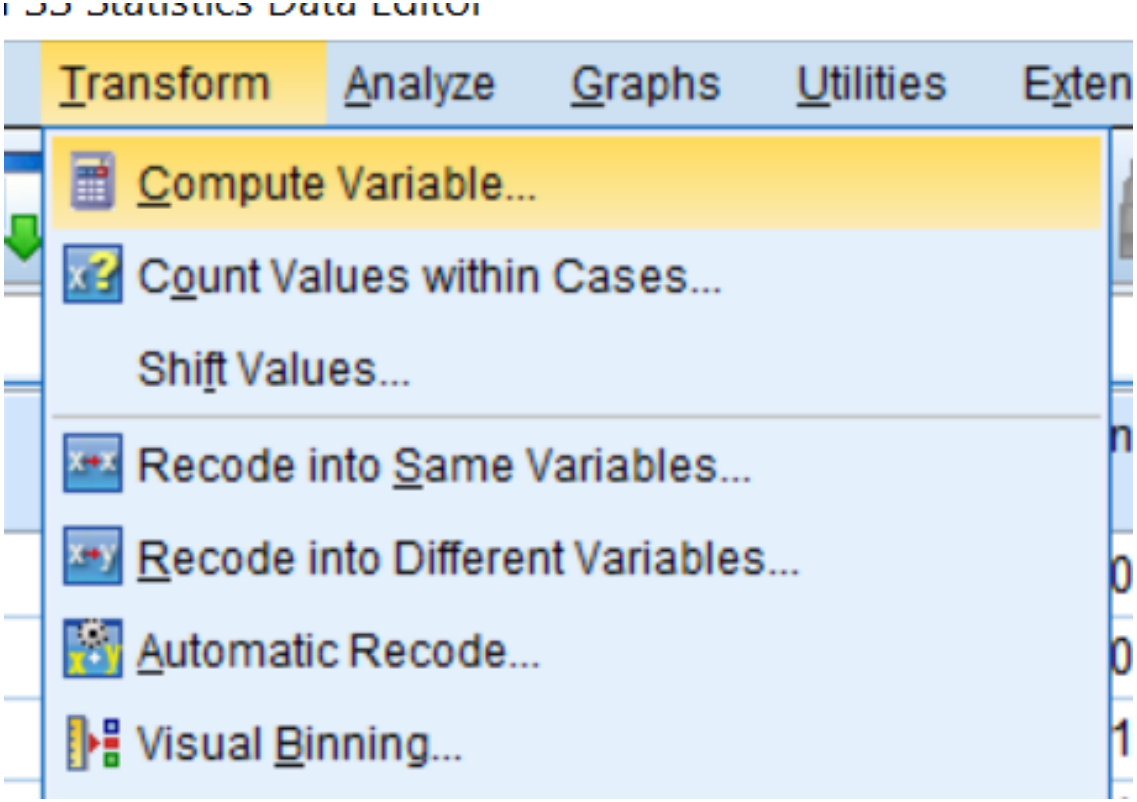
Correlations

		price	year
price	Pearson Correlation	1	.720**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	7.891E+10	32624714.38
	Covariance	158145979.2	65380.189
	N	500	500
year	Pearson Correlation	.720**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	32624714.38	25995.158
	Covariance	65380.189	52.095
	N	500	500

**. Correlation is significant at the 0.01 level (2-tailed).

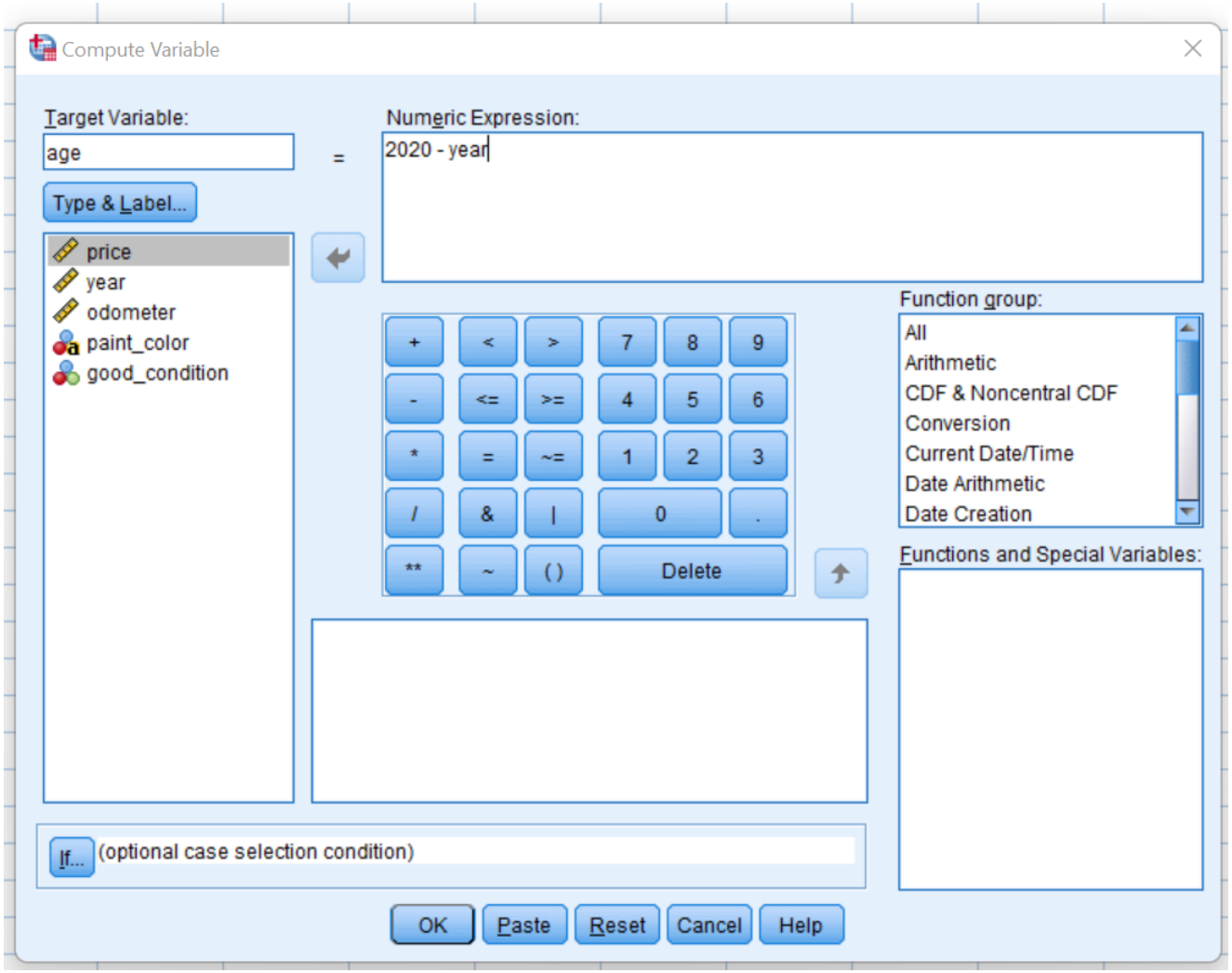
- Covariance is \$65,380.189. This indicates a positive relationship, but is otherwise not easily interpretable.
- Correlation is 0.720. This indicates a strong positive linear relationship (because the correlation can be at most 1).

Q6: Computing a New Variable, where $age = 2020 - year$



Q6: Computing a New Variable, where $age = 2020 - year$

For example, if year is 2019, age is 1. If year is 2018, age is 2.



Q6: Computing a New Variable

A new column appears in the dataset. Check that the rows make sense! Remember that the advertisements were from 2020.

	price	year	odometer	paint_color	good_condition	age
1	3000	1979	150000	orange	0	41.00
2	3200	1993	190000	silver	0	27.00
3	37800	2017	52653	red	1	3.00
4	16450	2016	49500	silver	1	4.00
5	7500	2008	240000	grey	1	12.00
6	19900	2016	105905	red	1	4.00
7	31995	2016	75280	black	1	4.00
8	11140	2006	114038	red	1	14.00
9	18500	2013	119347	black	1	7.00
10	3900	2003	160000	grey	0	17.00
11	5000	2005	205315	white	0	15.00
12	3800	1992	157600	blue	0	28.00
13	2250	1998	290000	white	0	22.00
14	1700	1993	212003	green	0	27.00
15	20500	2015	102200	red	1	5.00
16	38999	2015	53398	black	1	5.00
17	12500	2005	152000	red	0	15.00
18	1000	2018	100000	green	1	2.00
19	4600	2012	120000	white	0	8.00
20	15998	2017	84000	white	1	3.00
21	19950	2011	146029	red	0	9.00
22	50500	2020	6577	white	0	.00

Q7: Covariance and Correlation with Price and Age

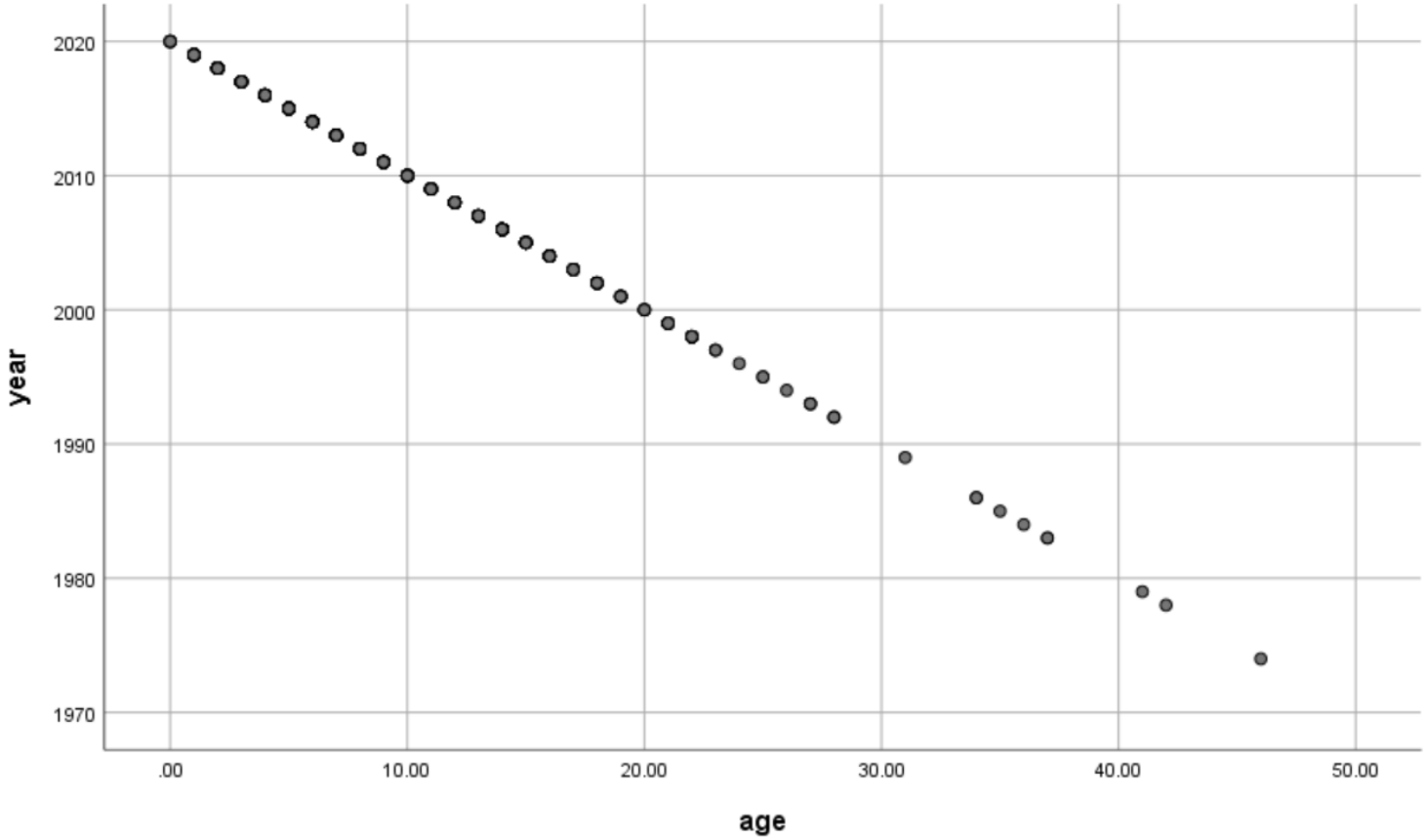
Correlations

		price	age
price	Pearson Correlation	1	-.720**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	7.891E+10	-32624714.4
	Covariance	158145979.2	-65380.189
	N	500	500
age	Pearson Correlation	-.720**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	-32624714.4	25995.158
	Covariance	-65380.189	52.095
	N	500	500

**. Correlation is significant at the 0.01 level (2-tailed).

- Covariance is -$65,380.189. This indicates a negative relationship, but is otherwise not easily interpretable. It has the same as with *Year* instead of *Age* apart from the negative sign.
- Correlation is -0.720. This indicates a strong negative linear relationship (because the correlation cannot be lower than -1). It has the same as with *Year* instead of *Age* apart from the negative sign.

Q8: Relationship Between Year and Age



- Perfect negative linear relationship!
- This is because $year = 2020 - age$ for every observation.

Q8: Relationship Between Year and Age

Correlations

		year	age
year	Pearson Correlation	1	-1.000**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	25995.158	-25995.158
	Covariance	52.095	-52.095
	N	500	500
age	Pearson Correlation	-1.000**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	-25995.158	25995.158
	Covariance	-52.095	52.095
	N	500	500

**. Correlation is significant at the 0.01 level (2-tailed).

- Correlation exactly -1. This is a perfect negative linear relationship.
- Both year and age have the same variance (52.095). The covariance is the negative of the variance (-52.095).

Bonus Questions from the Exam

Q1:

- The correlation is -0.7233.
- Any answer in the interval $[-0.724, -0.720]$ was accepted.
- Interpretation (not asked): There is a strong negative linear relationship between the truck's mileage and its advertised price.

Q2:

- Answer: *There is a negative linear relationship between the age of an F-150 and its price*.
- Explanation: The correlation between price and year was positive, but a higher year means a lower age, so there is a negative linear relationship between age and price.
- It wasn't necessary to create the *age* variable, but it helped to check.
- Other answer options were either stating a positive relationship between age and price, or no relationship, which we know is wrong!

.