Please write your name and netID:

Name:Mina Mohammadi NetID:mm9494

DS-UA 111: Homework Two

In pursuit of happiness (and data science, but I repeat myself)!

This homework is due Monday, March 9 by 8:00p. Late homework will be graded down, no exceptions. Each improperly formatted question will be marked down 20%. Note that the course academic honesty policy applies to every homework, including this one. Some of the questions refer to articles, which you can find by clicking on the links provided. This homework is worth 37 points (one point per sub-question).

Instructions

Please complete your answers in the spaces provided. They will be either cells for code or Markdown, and we will make it clear within each question how to reply.

The submission process for this and all assignments is explained in separate documentation from Lecture 2.2. Make sure to submit your assignment in both NB grader and GradeScope as described in Lecture 2.2.

Question 1

A researcher is interested in finding out what makes some people happier than others. In the following questions, please respond to prompts related to several steps of this research project.

(a) The researcher suspects that those who have a satisfying job, good health, and a robust social network are more likely to be happy compared to people who do not have those things. What is the dependent variable in this case?

Type the letter of the option that represents the dependent variable(s):
A. A satisfying job
B. Good health
C. Happiness
D. A robust social network
E. None of the above

In [3]:
```
q1a_answer = 'C'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q1a_answer' in locals()
assert q1a_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(b) In order to collect any data to illuminate this theory, the researcher must be specific about how to measure these concepts. They decide to measure "health" in terms of having a long life span. What step in the measurement process are they taking?

Type the letter of the option that reflects your answer:
A. Hypothesis formation
B. Estimation
C. Operationalization
D. Conceptualization
E. None of the above

In [4]:
```
q1b_answer = 'C'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q1b_answer' in locals()
assert q1b_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(c) The researcher has doubts that longevity is truly capturing health. What kind of error is the researcher worried about committing?

Type the letter of the option that reflects your answer:
A. Validity
B. Selection bias
C. Exclusion
D. Random
E. A & B

In [5]:
```
q1c_answer = 'E'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q1c_answer' in locals()
assert q1c_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(d) The researcher suspects that not everyone who has a satisfying job is happy, but that everyone who is happy must have a satisfying job. The researcher is describing what kind of conditional causal relationship?

Type the letter of the option that reflects your answer:
A. Having a satisfying job is a necessary condition for being happy
B. Having a satisfying job is a sufficient condition for being happy
C. Being happy is a necessary condition for having a satisfying job
D. Being happy is a sufficient condition for having a satisfying job
E. Having a satisfying job is both necessary and sufficient for being happy
F. Having a satisfying job is neither necessary nor sufficient for being happy
F. Being happy is both necessary and sufficient for having a job
G. Being happy is neither necessary nor sufficient for having a job
H. Make this stop

In [6]:
```
q1d_answer = 'A'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q1d_answer' in locals()
assert q1d_answer.upper() in ['A','B','C','D','E', 'F', 'G', 'H']
### END PUBLIC TESTS
```

(e) The researcher decides to test their theory by evaluating people who are happy to see what factors they have in common. This is an example of:

Type the letter of the option that reflects your answer
A. A randomized, controlled experiment
B. Selecting on the dependent variable
C. A natural experiment
D. Selecting on the independent variable
E. A and B
F. C and D
G. None of the above

In [7]:
```
q1e_answer = 'B'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q1e_answer' in locals()
assert q1e_answer.upper() in ['A','B','C','D','E', 'F', 'G']
### END PUBLIC TESTS
```

Question 2

Happiness and well-being has been the subject of much research. The following questions refer to the study described in this article, '[Good Genes are Nice, But Joy is Better (https://news.harvard.edu/gazette/story/2017/04/over-nearly-80-years-harvard-study-has-been-showing-how-to-live-a-healthy-and-happy-life/).](https://news.harvard.edu/gazette/story/2017/04/over-nearly-80-years-harvard-study-has-been-showing-how-to-live-a-healthy-and-happy-life/)'

(a) Without commenting on the study itself, what word is misused in the subtitle of the article?

Proved

(b) When the study began more than 80 years ago, the subjects were only male. Critics of the original study might argue that this is an example of exclusion, or invisibility, bias against non-men (and one could presumably point out other exlusion biases related to race, class, and more). Suppose a defender of the study says we can still learn from it because there's no reason to believe that the factors that make men happy are different for other genders. This defender is making an argument based on what?

Type the letter of the option that reflects your answer:
A. A lack of survivorship bias means the results are reliable
B. The fact that the conceptualization of happiness is robust
C. An assumption of orthogonality between happiness factors and gender
D. The importance of future work considering the results for other genders
E. None of the above

In [8]:
```
q2b_answer = 'C'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q2b_answer' in locals()
assert q2b_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(c) Another critic of the study argues that a major finding, that "happy relationships have a powerful influence on health", is problematic because of the possibility that health also contributes to happy relationships. What type of causal relationship is this critic invoking?

Sufficent but not necessary

(d) The article reports on this finding: "people who had happy marriages in their 80s reported that their moods didn't suffer even on the days when they had more physical pain. Those who had unhappy marriages felt both more emotional and physical pain." A critic points out a possibility that there might be an underlying third variable that explains both unhappy marriages and pain. This critic is describing the possibility of:

Type the letter of the option that reflects your answer:
A. Selection on the dependent variable
B. Survivorship bias
C. Selection bias
D. A confounder
E. None of the above

In [9]:
```
q2d_answer = 'D'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q2d_answer' in locals()
assert q2d_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(e) How is "emotional pain" conceptualized, according to the quote in question 2d?

Type the letter of the option that reflects your answer:
A. Overall well-being
B. Self-reported mood
C. Incidences of depression and/or anxiety
D. All of the above
E. None of the above

In [10]:
```
q2e_answer = 'A'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q2e_answer' in locals()
assert q2e_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(f) How is "emotional pain" operationalized, according to the quote in question 2d?

Type the letter of the option that reflects your answer:
A. Structured interviews
B. Random sampling
C. Surveys
D. Not enough information
E. Too much information

In [11]:
```
q2f_answer = 'C'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q2f_answer' in locals()
assert q2f_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(g) We've hated on this research a lot. What's one thing you think is a strength of this research?

It compiles data from a generally underrepresented group

Question 3: Getting to know a dataset

The researcher decides that rather than collect a whole bunch of data themselves or rely on the Harvard study, they are going to start with some pre-collected data and do their own analyses. They decide to use data from the 2019 World Happiness Report (https://worldhappiness.report) for their initial research.

(a) There is a dataset, `happiness.csv`, that lives in our JupyterHub shared folder. We have provided starter code that gets the full *path* of this file and stores it in `csv_fpath`. Use Pandas' `read_csv()` function to load it and store its contents into a DataFrame with the very creative name `data`.

In [12]:
```
import pandas as pd
from os.path import expanduser
csv_fpath = expanduser("~/shared/happiness.csv")
data = pd.read_csv(csv_fpath)
```

In [4]:
```
### BEGIN PUBLIC TESTS
assert 'data' in locals()
assert data.size == 44304
### END PUBLIC TESTS
```

```
---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call last)
<ipython-input-4-af575cd22254> in <module>
      1 ### BEGIN PUBLIC TESTS
----> 2 assert 'data' in locals()
      3 assert data.size == 44304
      4 ### END PUBLIC TESTS

AssertionError:
```

(b) Inspect the first five rows of the data using the `head` command to make sure it imported properly.

In [13]:
```
data.head()
```

Out[13]:

	Country name	Year	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perception c corruptio
0	Afghanistan	2008	3.723590	7.168690	0.450662	50.799999	0.718114	0.177889	0.88168
1	Afghanistan	2009	4.401778	7.333790	0.552308	51.200001	0.678896	0.200178	0.85003
2	Afghanistan	2010	4.758381	7.386629	0.539075	51.599998	0.600127	0.134353	0.70676
3	Afghanistan	2011	3.831719	7.415019	0.521104	51.919998	0.495901	0.172137	0.73110
4	Afghanistan	2012	3.782938	7.517126	0.520637	52.240002	0.530935	0.244273	0.77562

5 rows × 26 columns

(c) How many rows and columns are there in this dataset? Assign the row number to a variable called `row_num` and the column number to a variable called `col_num`.

In [14]:
```
row_num = data.shape[0]
col_num = data.shape[1]
print(row_num)
print(col_num)
```

```
1704
26
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'row_num' in locals()
assert 'col_num' in locals()
### END PUBLIC TESTS
```

(d) The output for the above question (3c) tells us what about the dataset?

Type the letter of the option that reflects your answer:
A. There are 1704 variables
B. There are 1704 observations
C. There are 26 variables
D. There are 26 observations
E. A and D
F. B and C
G. None of the above

In [15]:
```
q3d_answer = 'F'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q3d_answer' in locals()
assert q3d_answer.upper() in ['A','B','C','D','E', 'F', 'G']
### END PUBLIC TESTS
```

(e) Notice when we inspect the dataset using `head`, we can't see all the columns. Write the line of code that allows us to see exactly 26 columns, then use `.head()` to display them.

In [17]:
```
pd.set_option("display.max.columns", None)
data.head()
```

Out[17]:

	Country name	Year	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perception c corruptio
0	Afghanistan	2008	3.723590	7.168690	0.450662	50.799999	0.718114	0.177889	0.88168
1	Afghanistan	2009	4.401778	7.333790	0.552308	51.200001	0.678896	0.200178	0.85003
2	Afghanistan	2010	4.758381	7.386629	0.539075	51.599998	0.600127	0.134353	0.70676
3	Afghanistan	2011	3.831719	7.415019	0.521104	51.919998	0.495901	0.172137	0.73110
4	Afghanistan	2012	3.782938	7.517126	0.520637	52.240002	0.530935	0.244273	0.77562

(f) The variable "Life Ladder" is the variable representing happiness in this study. According to the codebook associated with the study, how is "Life Ladder" conceptualized?

Life Ladder is conceptualied as happiness and is measured by averageing the respondant answers to the Centril ladder survey

(g) Comment on one strength and one weakness of this conceptualization.

One strenth is that it operationalizes and polls many people easily but one con is that it relies on own people's interpretations of their happiness/self-reported data

(h) Life Ladder is based on self-reported data. Some may be critical that this might impose selection bias; specifically, that people might be more inclined to report they are happier than they are when speaking to strangers. If this is the case, in what direction would this bias the results?

Please write the letter that reflects your response:
A. It would make people seem happier than they really are
B. It would make people seem less happy than they really are
C. It would make people seem more extreme -- either much happier or less happier -- than they really are
D. It would make people seem less extreme -- most replies would be towards the middle -- than they really are
E. It would not have an effect on the results

In [18]:
```
q3h_answer = 'A'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q3h_answer' in locals()
assert q3h_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

Question 4: Initial analysis

The researcher now begins to explore initial trends in the dataset.

(a) In data science, it's common to rename variables we expect to use a lot in order to make them easier to reference. Rename the following variables exactly as indicated:
"Life Ladder" as "ladder"
"Log GDP per capita" as "gdp"
"Perceptions of corruption" as "corruption"

In [21]:
```
data= data.rename(columns={"Life Ladder": "ladder", "Log GDP per capita": "gdp", "Perceptions of corruption": "corruption"})
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'data' in locals()
### END PUBLIC TESTS
```

(b) Write a line of code to find the mean of your newly renamed variable `ladder`, and assign it to the variable `ladder_mean`.

In [22]:
```
ladder_mean = data['ladder'].mean()
print(ladder_mean)
```

```
5.437155035744718
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'ladder_mean' in locals()
### END PUBLIC TESTS
```

(c) Notice so far we are working with all countries and all years in the dataset. Use the `describe` command to write a line of code to find the most recent year in the dataset.

In [23]:
```
data["Year"].describe()
```

Out[23]:
```
count    1704.000000
mean     2012.332160
std         3.688072
min      2005.000000
25%      2009.000000
50%      2012.000000
75%      2015.000000
max      2018.000000
Name: Year, dtype: float64
```

(d) Write one line of code to create a new dataframe containing only data from 2018 and call it `data18`. Write a second line of code to inspect the first five rows of `data18`.

In [24]:
```
data18 = data[data.Year == 2018].copy()
data18.head()
```

Out[24]:

	Country name	Year	ladder	gdp	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	corruptio
10	Afghanistan	2018	2.694303	7.494588	0.507516	52.599998	0.373536	-0.084888	0.92760
21	Albania	2018	5.004403	9.412399	0.683592	68.699997	0.824212	0.005385	0.89912
28	Algeria	2018	5.043086	9.557952	0.798651	65.900002	0.583381	-0.172413	0.75870
45	Argentina	2018	5.792797	9.809972	0.899912	68.800003	0.845895	-0.206937	0.85525
58	Armenia	2018	5.062449	9.119424	0.814449	66.900002	0.807644	-0.149109	0.67682

(e) Find the least happy country in 2018 by writing one line of code that sorts the data by `ladder`, uses the `ascending` argument, and shows the first five rows, where the first row is the least happy country in 2018 according to the `ladder` measure.

In [25]:
```
data.sort_values('ladder', ascending = True).head(5)
```

Out[25]:

	Country name	Year	ladder	gdp	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	corrup
9	Afghanistan	2017	2.661718	7.497755	0.490880	52.799999	0.427011	-0.112198	0.954
1458	Syria	2013	2.687553	8.307246	0.585450	58.759998	0.454883	0.224178	0.663
274	Central African Republic	2016	2.693061	6.465948	0.290184	44.900002	0.624057	0.053440	0.859
10	Afghanistan	2018	2.694303	7.494588	0.507516	52.599998	0.373536	-0.084888	0.927
864	Liberia	2015	2.701591	6.665998	0.637666	53.700001	0.671431	-0.014139	0.902

(f) Find the happiest country in 2018 by writing one line of code that sorts the data by `ladder` and shows the first five rows, where the first row is the happiest country in 2018 according to the `ladder` measure.

In [26]:
```
data.sort_values('ladder', ascending = False).head(5)
```

Out[26]:

	Country name	Year	ladder	gdp	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	corruptio
397	Denmark	2005	8.018934	10.704770	0.972372	69.599998	0.971135	NaN	0.23652
399	Denmark	2008	7.970892	10.733475	0.953912	70.080002	0.969788	0.267413	0.24750
495	Finland	2018	7.858107	10.636060	0.962155	71.900002	0.937807	-0.131735	0.19860
398	Denmark	2007	7.834233	10.744484	0.954201	69.919998	0.932086	0.235316	0.20600
494	Finland	2017	7.788252	10.611172	0.963826	71.800003	0.962199	-0.005696	0.19241

(g) Find the mean, standard deviation, max, and min of `ladder` for countries in 2018 using `describe`.

In [27]:
```
data['ladder'].describe()
```

Out[27]:
```
count    1704.000000
mean        5.437155
std         1.121149
min         2.661718
25%         4.610970
50%         5.339557
75%         6.273522
max         8.018934
Name: ladder, dtype: float64
```

(h) What does the standard deviation tell us about the data in a variable?

Type the letter that reflects your answer:
A. Spread-outness
B. Reliability
C. Validity
D. Central tendency
E. All of the above
F. None of the above

In [28]:
```
q4h_answer = 'E'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q4h_answer' in locals()
assert q4h_answer.upper() in ['A','B','C','D','E','F']
### END PUBLIC TESTS
```

(i) What does the 25% output for `ladder` in Question 4g mean, in words?

This is the first quartile, 25% of the ladder values are less than or equal to 4.61.

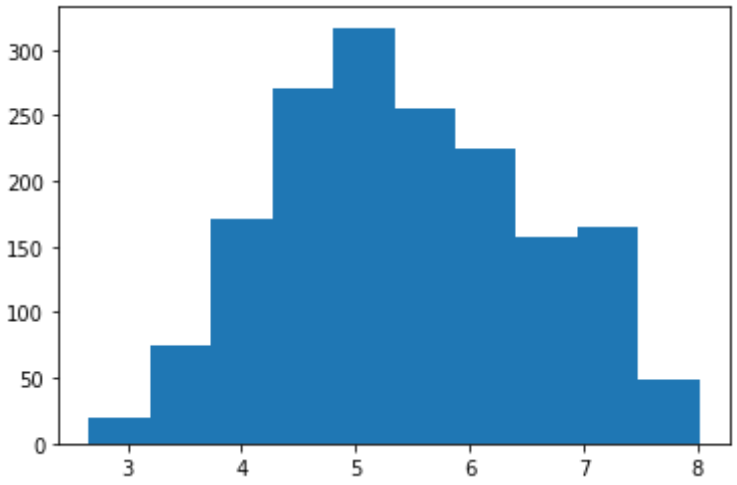
Question 5: Visualization

The researcher now explores trends in the dataset using visualizations.

(a) We've been discussing the numeric distributions of `ladder`. Histograms are a great tool for understanding distributions of a variable of interest. Write one line of code to generate a histogram of `ladder` for 2018. Do not worry about specifying bins or labels for now.

In [29]:
```
import matplotlib.pyplot as plt
plt.hist(data.ladder)
```

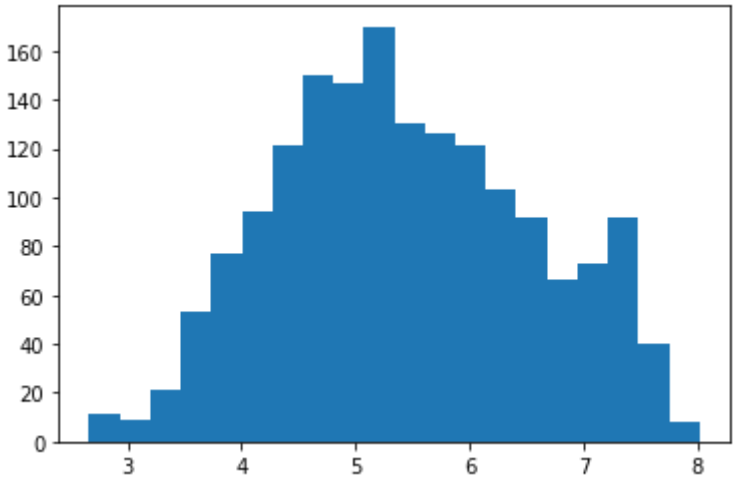
Out[29]:
```
(array([ 20.,  74., 171., 271., 317., 256., 224., 158., 165.,  48.]),
 array([2.66171813, 3.19743974, 3.73316135, 4.26888297, 4.80460458,
        5.34032619, 5.8760478 , 6.41176941, 6.94749103, 7.48321264,
        8.01893425]),
 <a list of 10 Patch objects>)
```



(b) To better understand a distribution, we may want to experiment with binning. Write one line of code to create a histogram that is otherwise the same, but set `bins` = 20. Observe the difference between this and the above.

In [30]:
```
plt.hist(data.ladder, bins=20)
```

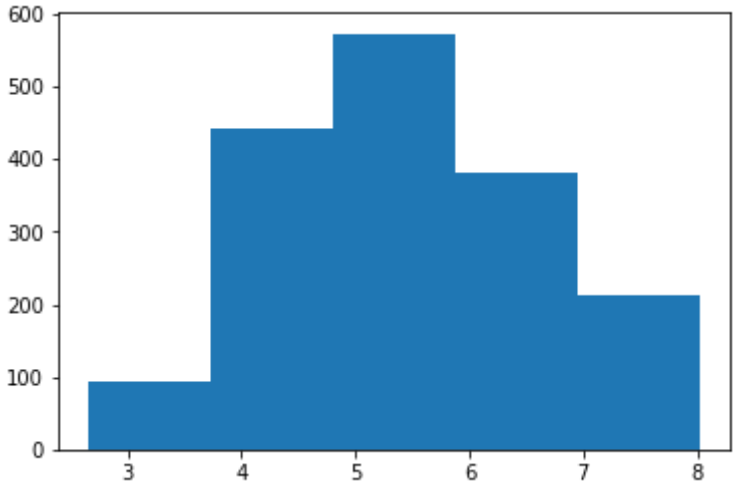
Out[30]:
```
(array([ 11.,   9.,  21.,  53.,  77.,  94., 121., 150., 147., 170., 130.,
        126., 121., 103.,  92.,  66.,  73.,  92.,  40.,   8.]),
 array([2.66171813, 2.92957894, 3.19743974, 3.46530055, 3.73316135,
        4.00102216, 4.26888297, 4.53674377, 4.80460458, 5.07246538,
        5.34032619, 5.608187  , 5.8760478 , 6.14390861, 6.41176941,
        6.67963022, 6.94749103, 7.21535183, 7.48321264, 7.75107344,
        8.01893425]),
 <a list of 20 Patch objects>)
```



(c) Now write one line of code to create a histogram that is otherwise the same, but set bins = 5. Observe the difference between this and the above.

In [31]:
```
plt.hist(data.ladder, bins=5)
```

Out[31]:
```
(array([ 94., 442., 573., 382., 213.]),
 array([2.66171813, 3.73316135, 4.80460458, 5.8760478 , 6.94749103,
        8.01893425]),
 <a list of 5 Patch objects>)
```



(d) We've now explored the distribution of `ladder' at a variety of levels of granularity. Based on these observations, using our beloved "ocular method", how do the data in this variable seem to be distributed?

Type the letter of the answer that reflects your response.
A. Respondents generally report similar levels of happiness
B. Respondents generally report either really happy or really low levels of happiness
C. Respondents generally report pretty central levels of happiness that are all close to the mean
D. Respondents generally report central levels of happiness, though more tend to be just above or below the mean rather than on it.
E. We can't tell any of this from this histogram

In [32]:
```
q5d_answer = 'D'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q5d_answer' in locals()
assert q5d_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(e) The researcher wants to understand changes in happiness over time in the US. Create a new variable called `dataus` that contains just observations of the US, making sure to use the `.copy()` function so that `dataus` is a separate DataFrame from (rather than a view of) `data`. Call `display` on the first five rows to make sure it worked as expected.

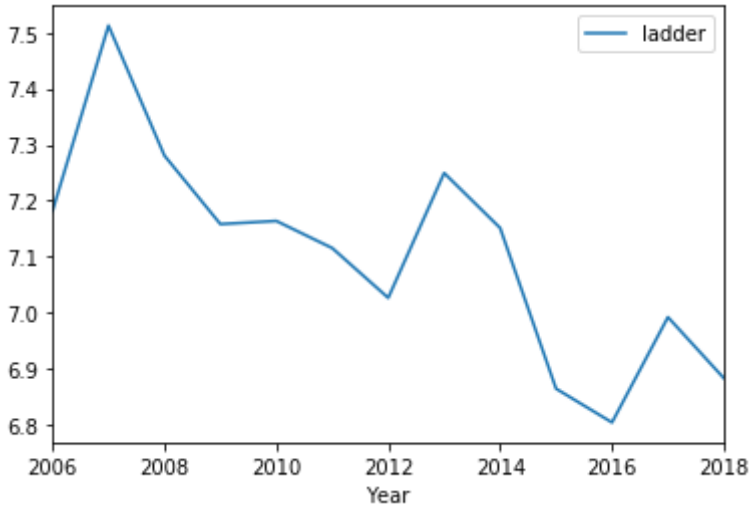
In [34]:
```
data = data.rename(columns={"Country name": "Country"})
dataus = data[data.Country == 'United States'].copy()
display(dataus.head(5))
```

	Country	Year	ladder	gdp	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	corruptio
1604	United States	2006	7.181794	10.831686	0.964572	68.059998	0.911496	NaN	0.60030
1605	United States	2007	7.512688	10.839805	NaN	68.220001	0.871904	0.188272	0.63303
1606	United States	2008	7.280386	10.827426	0.952587	68.379997	0.877956	0.246029	0.66849
1607	United States	2009	7.158032	10.790511	0.911794	68.540001	0.830684	0.192269	0.66539
1608	United States	2010	7.163616	10.807183	0.926159	68.699997	0.828044	0.235522	0.68958

(f) Line graphs are useful for visualizing trends over time. Use `matplotlib` to generate a line graph of levels of happiness in the United States over all years available for the US. For this line graph, specify that the x-axis shoud be years.

In [35]:
```
dataus.plot.line(x='Year', y='ladder')
```

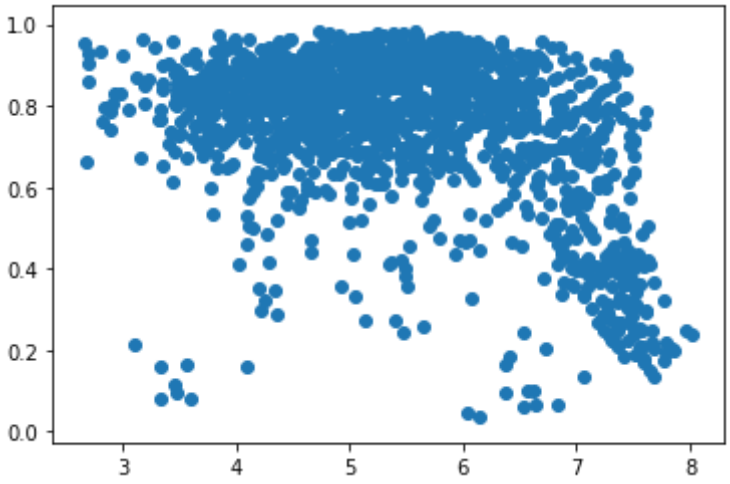
Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0f84805410>



(g) Scatterplots are useful for inspecting associations between variables. Create a scatterplot that compares `ladder` against `corruption` for 2018, where `ladder` is on the x-axis.

In [36]:
```
plt.scatter(data.ladder,data.corruption)
```

Out[36]: <matplotlib.collections.PathCollection at 0x7f0f8470bfd0>



(h) Based on the scatterplot in the previous question, what is the likely association between corruption and happiness? (E.g., as corruption goes down, what seems to be happening to happiness?)

Type the letter of the response that reflects your answer:

A. Strongly positive
B. Weakly positive
C. No relationship
D. Weakly negative
E. Strongly negative

In [37]:
```
q5h_answer = 'D'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q5h_answer' in locals()
assert q5h_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

End of Homework

Please remember to submit it correctly -- as shown in Lecture 2.2 and practiced in Lab 0!

That means Jupyterhub and Gradescope!