BUILDING A LINEAR REGRESSION MODEL

USING PYTHON AND PANDAS

DEPRESSION & SOCIAL MEDIA

According to the World Health Organization, depression is the leading cause of disability worldwide, and is a major contributor to the overall global burden of disease. Due to access barriers to treatment depression also is often under-reported.

With also a good friend suffering from depression I was interested in which factors can especially trigger, worsen and affect the state of mental health.
I was aware of the ongoing research on the topic and that there actually is a link between social media and depression, that especially affects young girls.

What I was particularly interested in was, how social media affects severe depression. As an estimator for that I chose antidepressant consumptions since the usual procedure would be to first try a therapy approach before putting a patient under medication.

Is there a positive linear relationship between social media and antidepressant consumption?

DATA SOURCES

The data on the consumption of antidepressant was taken from a dataset on the consumption of different medication in the pharmaceutical market worldwide.
That dataset contains information on antidepressants for 30 countries worldwide by year from 2000 until 2018, although the data for the the latter is incomplete.*
The consumption is given by *Defined daily dosage per 1 000 inhabitants per day.*

For the Social Media Data for the Year 2019 a website with relevant data has been scraped.**
This website contains relevant information about 22 European countries.
The social media use is given by *Percentage of Active social media penetration.*

- https://stats.oecd.org/index.aspx?queryid=30123#
- ** https://www.statista.com/statistics/295660/active-social-media-penetration-in-european-countries/

DATA CLEANING & MANIPULATION

The data on the medication consumption was provided as a CSV file and then loaded as a Pandas Dataframe in Python:

```
# Importing libraries
import pandas as pd

# Read in dataset
df_pharma = pd.read_csv('HEALTH_PHMC_08092019205322815.csv')
```

In the next step, I filtered the dataset to only the antidepressant consumption and dropped columns that were irrelevant for the analysis:

```
# Drop columns (Measure, UNIT, VAR only have 1 value; Flags and Flag Code not relevant, YEA duplicate)
df_pharma_N06 = df_pharma_N06.drop(columns = ['Measure', 'Flag Codes', 'UNIT', 'VAR','YEA','Flags'])

# Reset index
df_pharma_N06 = df_pharma_N06.reset_index(drop=True)
```

The dataset looks now like this:

	Variable	COU	Country	Year	Value
0	N06A-Antidepressants	AUS	Australia	2000	45.4
1	N06A-Antidepressants	AUS	Australia	2001	53.2
2	N06A-Antidepressants	AUS	Australia	2002	54.7
3	N06A-Antidepressants	AUS	Australia	2003	58.5
4	N06A-Antidepressants	AUS	Australia	2004	63.4

WEBSCRAPING

Using Beautiful Soup, I scraped the data from the statista.com Website.

```
import requests
from bs4 import BeautifulSoup

url = 'https://www.statista.com/statistics/295660/active-social-media-penetration-in-european-countries/'
html = requests.get(url).content
 soup = BeautifulSoup(html, "lxml")
```

After the use of some Regex Code and dropping countries that were not present in the antidepressant dataset I had to do some more cleaning and drop duplicate values.

After what seemed like ages, I arrived at a dataframe that could finally be merged with the medication dataset.
Both datasets now include the same countries and the medication dataset was filtered to the Year 2017 (the most recent year with the most existing data).

```
combined_df = pd.merge(df_pharma_N06_2017,new_df_social[['Country','Percentage']], on='Country')
```

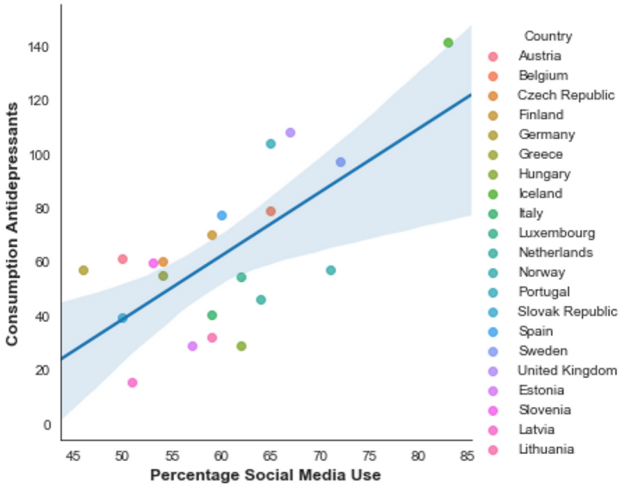
	Variable	COU	Country	Year	Value	Percentage
0	N06A-Antidepressants	AUT	Austria	2017	61.0	50
1	N06A-Antidepressants	BEL	Belgium	2017	78.8	65
2	N06A-Antidepressants	CZE	Czech Republic	2017	59.9	54
3	N06A-Antidepressants	FIN	Finland	2017	70.2	59
4	N06A-Antidepressants	DEU	Germany	2017	56.9	46
5	N06A-Antidepressants	GRC	Greece	2017	55.1	54

CHECKING FOR CORRELATIONS

First I plotted the two variables in a scatterplot to visually detect if there could be any correlations between social media use and antidepressant consumption:

```
# Importing libraries
import seaborn as sns
import matplotlib as plt
import matplotlib.pyplot as plt
%matplotlib inline

# Plot Scatterplot with regression line
g = sns.lmplot(x="Percentage", y="Value", hue="Country", data=combined_df, fit_reg=False)
sns.regplot(x="Percentage", y="Value", data=combined_df, scatter=False, ax=g.axes[0, 0])
plt.rcParams['figure.figsize']=9,7
plt.xlabel('Percentage Social Media Use',fontsize=12, fontweight="bold")
plt.ylabel('Consumption Antidepressants',fontsize=12, fontweight="bold")
plt.savefig('Linear Regression2.png',dpi=300)
plt.show()
```



AND YES! WE CAN DETECT A POSSIBLE LINEAR RELATIONSHIP

In the next step I calculated the correlation coefficient:

combined_df.corr()

	Year	Value	Percentage
Year	NaN	NaN	NaN
Value	NaN	1.000000	0.675057
Percentage	NaN	0.675057	1.000000

The calculated correlation coefficient of 0.657, which confirms that both variables seem to have a connection.

BUILDING A LINEAR REGRESSION MODEL

We fitted a positive linear regression line that is aiming at predicting antidepressants consumption with social media use for our model.

```
#Importing libraries
import scipy
from scipy import stats

X = combined_df['Percentage']
Y = combined_df['Value']

slope, intercept, r_value,p_value, std_err = stats.linregress(X, Y)

print(stats.linregress(X, Y))

LinregressResult(slope=2.3603219165262965, intercept=-79.55650383679584, rvalue=0.6750566803726034,
pvalue=0.0007867697578795408, stderr=0.591796999907399)
```

The best fit model resulted in the following formula.:

y = - 79.56 + 2.36 * x

With y being the dependent variable antidepressants consumption and x being our predictor variable social media use, we can interpret this such that **an increase of 2.36 % of social media use will lead to a 1 step increase in antidepressants consumption**.

REJECTING THE NULL HYPOTHESIS

The null hypothesis assumes that the use of social media . use has no impact on the antidepressant consumption.
We conduct a test to determine the p-value which can tell us if our correlation coefficient is statistically significantly different from 0. The value of p is the probability that the results occurred by chance. We obtained a **p-value** of **0.00079**, or in other words 0.079 %.
Therefore we can reject the null hypothesis at the 99% Confidence level and conclude that **our two variables are not independent of each other.**

MODEL FIT

How well can our model predict our model changes in the consumption of antidepressants?
We are calculating the R-squared value that evaluates the scatter of data points around the fitted regression line. It is a measure of fit that indicates the percentage of variance in the dependent variable explained by the independent variable.

```
print("R-squared: %f" % 0.6750566803726034**2)
R-squared: 0.455702
```

We obtain a **R-squared value of 0.456** on a 0 to 1 scale.
This is due to high variation of our data points around the line.

But what does this mean for our model?
It means that social media alone cannot explain antidepressant consumption. In order to improve our model, we would have to take other predictors into account as well.

In the **Future Adjustment** section we give suggestions for such predictors that can be included in future research

KEY FINDINGS

- SOCIAL MEDIA USE IN SELECTED EUROPEAN COUNTRIES HAS AN EFFECT ON ANTIDEPRESSANT CONSUMPTION
- IN OUR MODEL A 2.36 STEP INCREASE IN SOCIAL MEDIA USE ACCOUNTS FOR 1 STEP INCREASE IN ANTIDEPRESSANT CONSUMPTION
- HIGH VARIATION IN OUR MODEL - CANNOT BE SOLELY EXPLAINED BY SOCIAL MEDIA USE

FUTURE ADJUSTMENTS

In order to develop a well predicting model, we have to include several more predictor variables. For example hours of sunlight or the countries' GDP values could be of interest. Sunlight and darkness trigger the release of hormones in your brain. Exposure to sunlight is thought to increase the brain’s release of a hormone called serotonin.
Furthermore, if we really want to use the model as a worldwide predictor for sales, we would need to adjust the sample method. We either need to use cluster sampling for all countries or sample in a way that the chosen countries are representative for the world.