In [230…
```
import numpy as np
import pandas as pd
import altair as alt
import statsmodels.api as sm
import math
import pylab
import scipy.stats as stats
from statsmodels.stats.outliers_influence import variance_inflation_factor
import seaborn as sn
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
alt.data_transformers.disable_max_rows()
alt.renderers.enable('mimetype')
```

Out[230…
```
RendererRegistry.enable('mimetype')
```

# PSTAT 100 Final Project

## World Happiness Report Data

| Name | Interpretation | Type | Units of measurement |
|---|---|---|---|
| Country name | Name of the country | Categorical | None |
| year | Year | Numeric | Years |
| Life Ladder | Rating of how happy respondent is with their life on a scale of 1 - 10 | Numeric | None |
| Log GDP per capita | GDP per capita on log scale | Numeric | None |
| Social support | National average of the binary responses to the GWP question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?" | Categorical | None |

| Name | Interpretation | Type | Units of measurement |
|---|---|---|---|
| Healthy life expectancy at birth | Life expectancy a birth assuming health is good | Numeric | Years |
| Freedom to make life choices | National average of responses to the GWP question "Are you satisfied or dissatisfied with your freedo mto choose what you do with your life?" | Numeric | None |
| Generosity | National average of response to the GWP question "Have you donated money to a charity in the past month?" | Numeric | None |
| Perceptions of corruption | National average of binary responses to the GWP questions "Is corruption widespread thorughout thre government or not?" and "Is corruption widespread wihin businesses or not?" | Numeric | None |
| Positive affect | The average of three positive measures in GWP: laugh, enjoyment, and doing interesting things, measured by question responses | Numeric | None |
| Negative affect | The average of three negative measures in GWP: worry, sadness, and anger, measured by question reponses | Numeric | None |
| region | Geographic region of country | Categorical | None |
| sub-region | Geographic sub-region of country | Categorical | None |

The world happiness report offers a comprehensive view of global well-being through the evaluation of a diverse set of factors. The dataset is updated anually and provides valuable insights on the complex interplay of social, economic, and envirnomental elements. The data is pulled from the Gallup World Poll (GWP) which surverys at least 1,000 individuals from each country included in the dataset. Individuals were asked a range of questions relation to their financial well being, social life, environment, and life outlook.

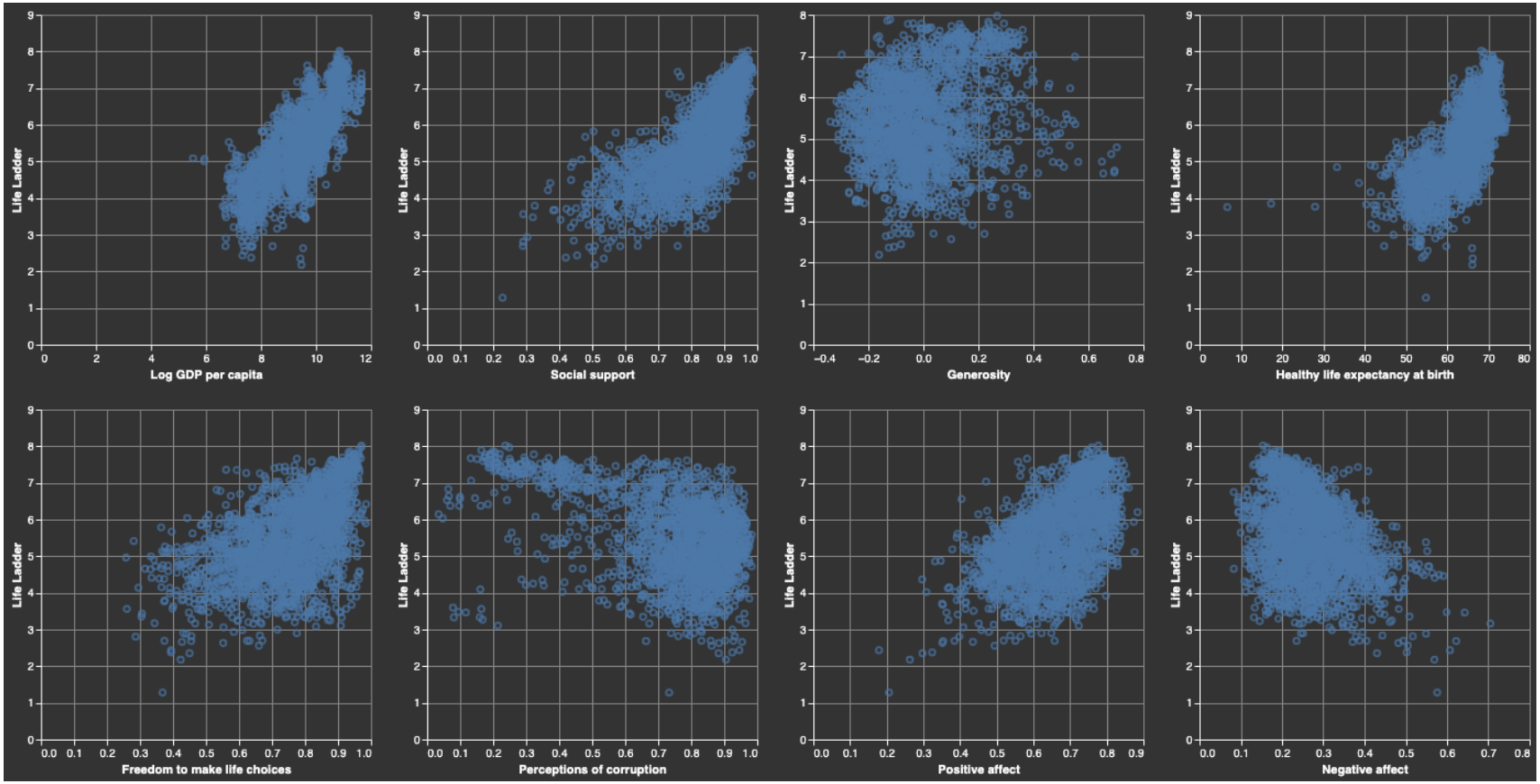
## Questions of Interest

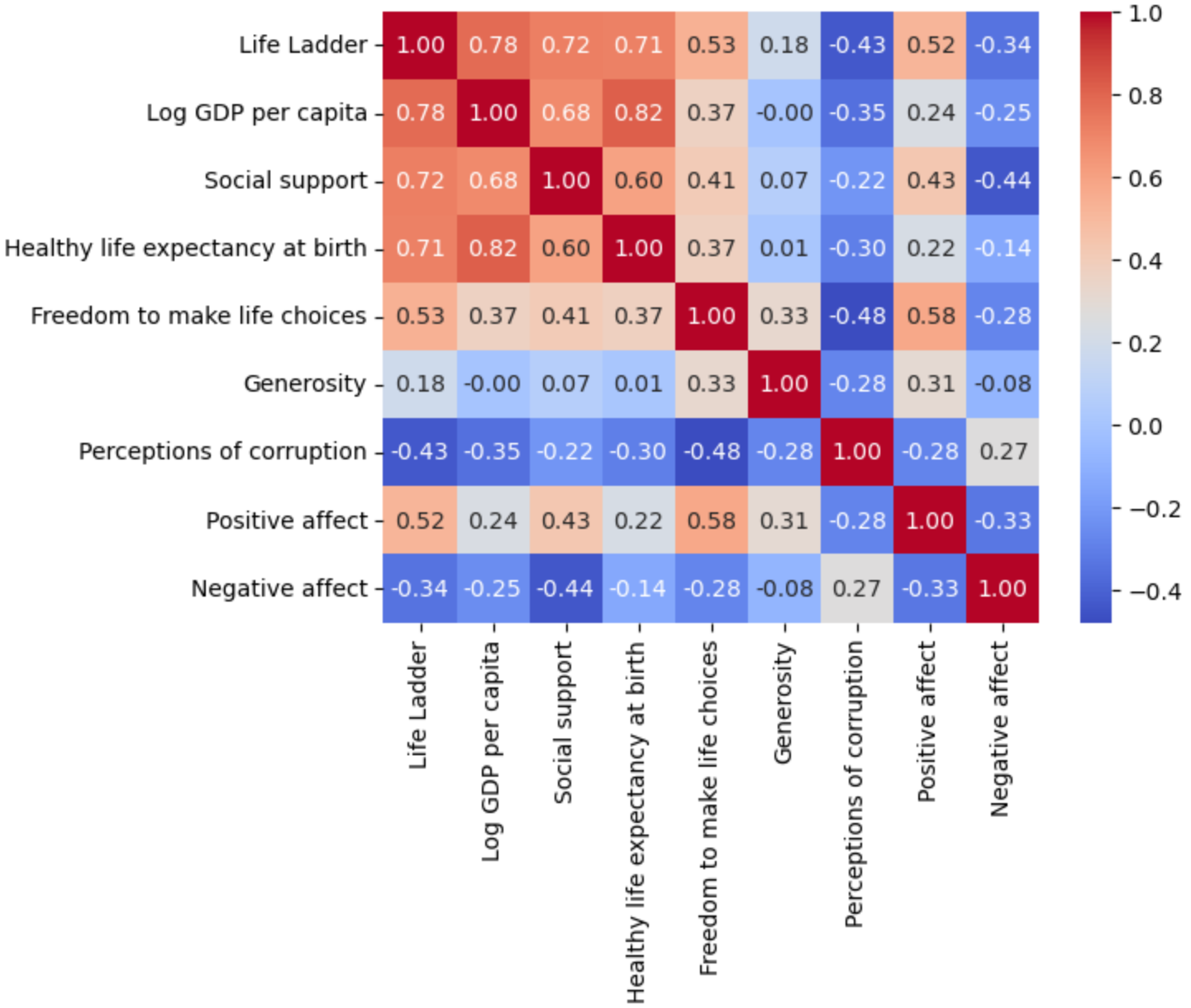
1. What is the relationship between life ladder and the other variables, is there any correlation?
2. Which variables are the most correlated with positive effect, and does money buy happiness?
3. Is there a discrepancy in life ladder across geographical regions and if so, why?
4. How did large scale world events affect happiness?
5. Given a set of features, are we able to predict happiness using linear regression?

## Question 1: What is the relationship between life ladder and the other variables, is there any correlation?

In [231…

```
display(life_ladder_scatter)
sn.heatmap(corr, annot = True, fmt = '.2f', cmap = 'coolwarm')
plt.show()
```

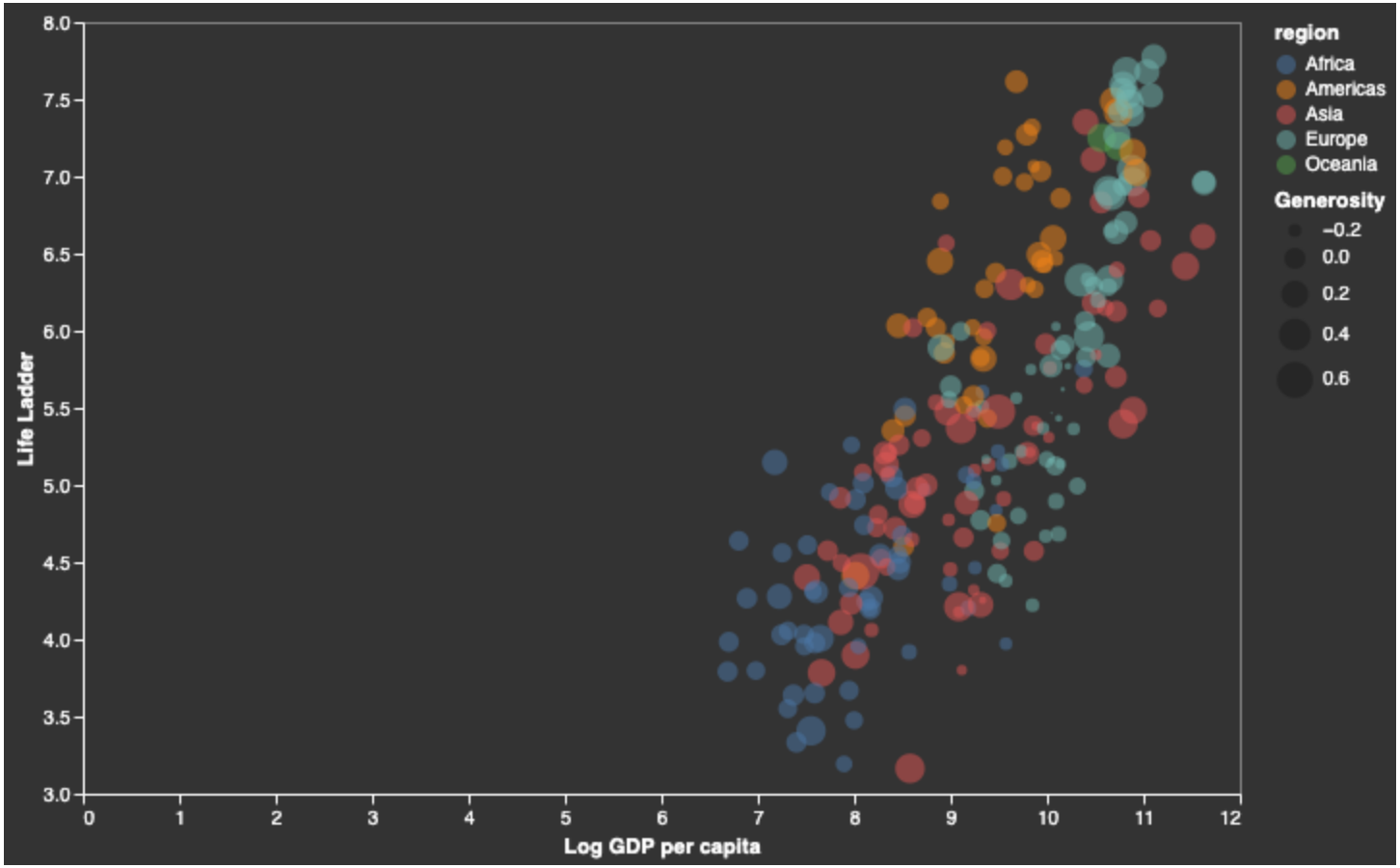





As we can see from the visuals, it seems as if most of the variables have a positive linear relationship with "Life Ladder" with the exception of "Generosity", which has almost no relationship. This is a surprising result and it seems to say that the more happy we are, the less we pay attention to the troubles of other people. There's a saying that misery loves company and this

seems to be the antithesis, that happy people are living in their own blissful world. Another conclusion we can draw from the visuals is that "Perceptions of corruption" and "Negative affect" have weak negative linear relationships, which makes sense bacause these quantify unhappiness. It seems that by a wide margin "Log GDP per capita", "Social support" and "Healthy life expectancy at birth" seem to have the strongest positve correlation with "Life Ladder" and we further explore this below.

In [232…
```
display(life_gdp)
```



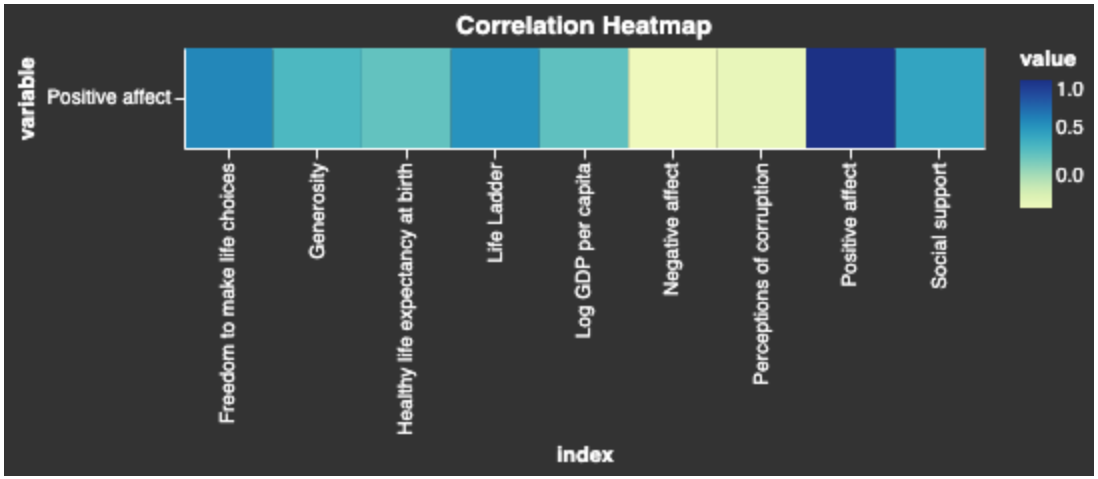

Our visual includes "Log GDP per capita" (x), "Healthy life expectancy at birth" (y), grouped by "region", and finally sized by "Generosity". Only data from 2009 and 2012 was examined, otherwise there would be too much data to visualize any trends. These years were chosen as they had the most median "Life Ladder" values, so it was likely that we would get a wide range of values (above and below the mean). Through this visual, we can see that there is a small cluster in the top right portion that is high on everything ("Generosity", "Log GDP per capita", and "Life Ladder"). Most of these are colored green which fall in

Europe. On the other hand, in the bottom left corner there is a cluster of small circles (likely a generosity below 0) which fall under Africa. It seems that the obvious conclusion is that wealthier regions tend to also have happier people living there. We saw above in the scatter plot and correlation matrix that there didn't seem to be a relationship between "Generosity" and "Life Ladder", however in the plot above we see that despite there is a pattern based on region. It seems that irrespective of "Life Ladder" and "Log GDP per capita", Asia seems to have evenly distributed generosity values while only wealthier and happier European countries seems to be generous. We're unable to explore this quirk further but it may be due to cultural differences between Eastern and Western societies.

## Question 2: Which variables are the most correlated with positive effect, and does money buy happiness?

In the previous question we analyzed our data with respect to "Life Ladder" which can be used as an overall measure of long term happiness. However, another approach we can take is analyzing these variables with respect to positive affect. Based on the questions posed by the GWP, positive affect seems to be a plausible stand in for short term happiness on a day to day basis.

In [233…
```
display(positive_affect_heatmap)
```



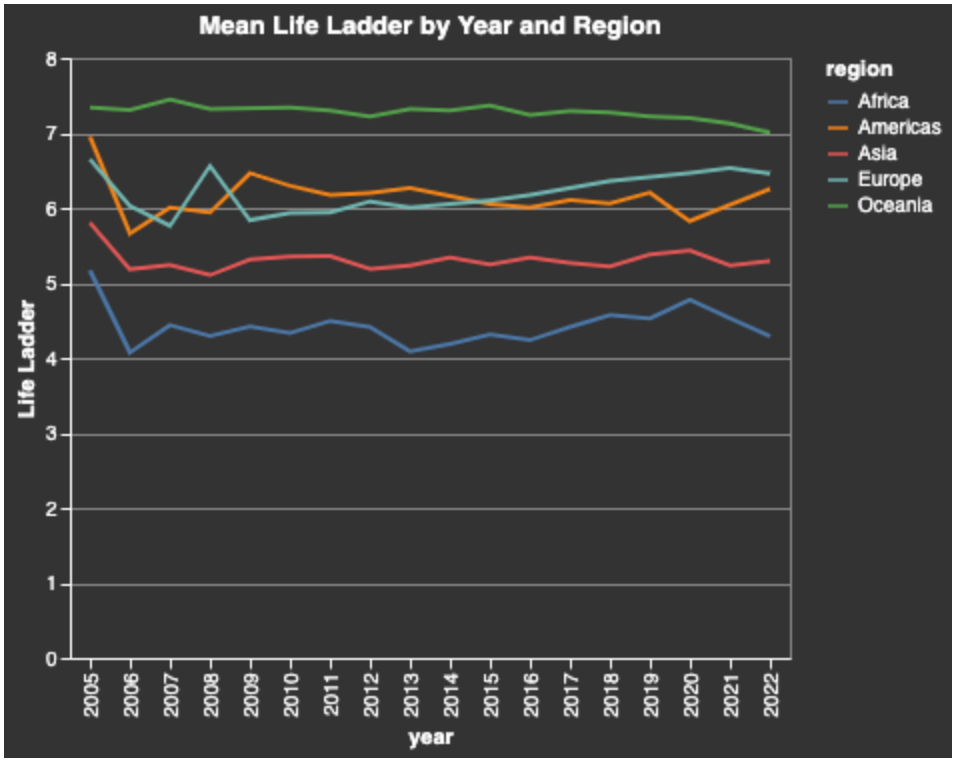

It seems that the two strongest correlations, excluding "Life Ladder", seem to be with "Freedom to make life choices" and "Social support". Based on the GWP questions used to quantify these two variables it seems that these aspects of happiness are very subjective, even more so than the other variables. These two variables seem to tap into the emotions of the

respondent; it makes sense that having friends/family you can depend on and feeling like you control the direction of your life directly imapcts your short term emotional state.

## Question 3: Is there a discrepancy in life ladder across geographical regions and if so, why?

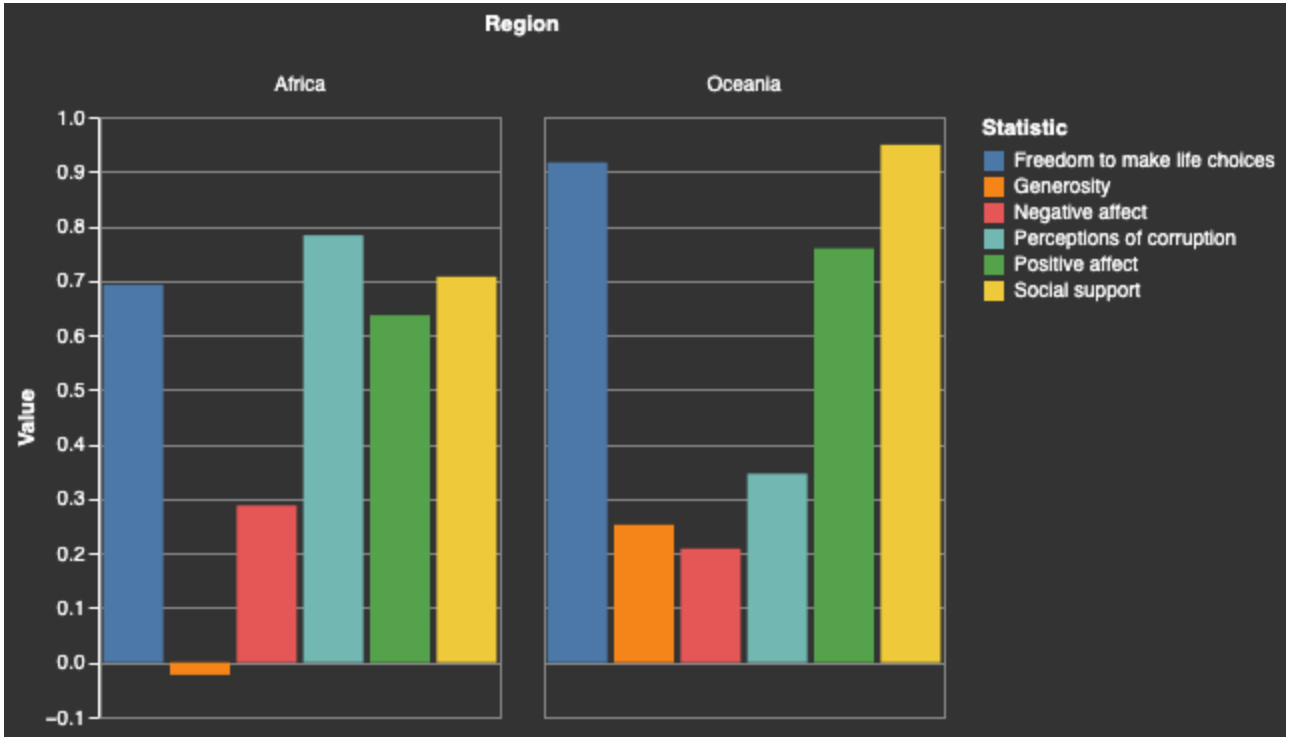
In [234…

```
display(life_ladder_region)
```




The first thing we did is to creater a "Life Ladder" vs "year" line graph and group by regions so we can see how happiness varies across countries and between years. It should come as no surprise that Oceania (which includes Australia/New Zealand) and Europe rank among the top regions in happiness while Africa ranks at the bottom. The obvious conclusion that we can draw from this is that first world countries enjoy a higher standard of living and are therefore happier. This is corroborated by our earlier analysis that shows that "Log GDP per capita" has the strongest correlation with "Life Ladder". Our next step is to compares how the other variables vary among the happiest region and the least happy region.

In [235…
```
display(most_least_happy)
```




Just as expected Oceania scores higher in all positive categories when compared to Africa and below we show that these differences in happiness are statistically significant.

In [236…
```
# Define the two groups
oceania = whr[whr['region'].isin(['Oceania'])]['Life Ladder']
africa = whr[whr['region'].isin(['Africa'])]['Life Ladder']

# Calculate the mean positive affect for each group
mean_oceania = oceania.mean()
mean_africa = africa.mean()

# Remove NaN values from the Europe, Africa, and Asia group
africa_cleaned = africa.dropna()

# Perform the t-test again with the cleaned data
t_stat, p_value = stats.ttest_ind(oceania, africa, equal_var = False)
```

```
# Display the new t-statistic and p-value
(t_stat, p_value)
```

Out[236… (81.05059504862174, 5.293039104143237e-153)

Using an alpha = 0.05 significance level, we can confidently conclude that the difference between the groups is statistically significant and such a small p-value really emphasizes just how big this discrepancy really is. One last thing I wanted to check out is the cultural makeup of the regions, i.e. how many countries make up each region.

In [237…
```
countries_per_region
```

Out[237…
```
region
Africa      49
Americas    27
Asia        47
Europe      40
Oceania      2
Name: Country name, dtype: int64
```

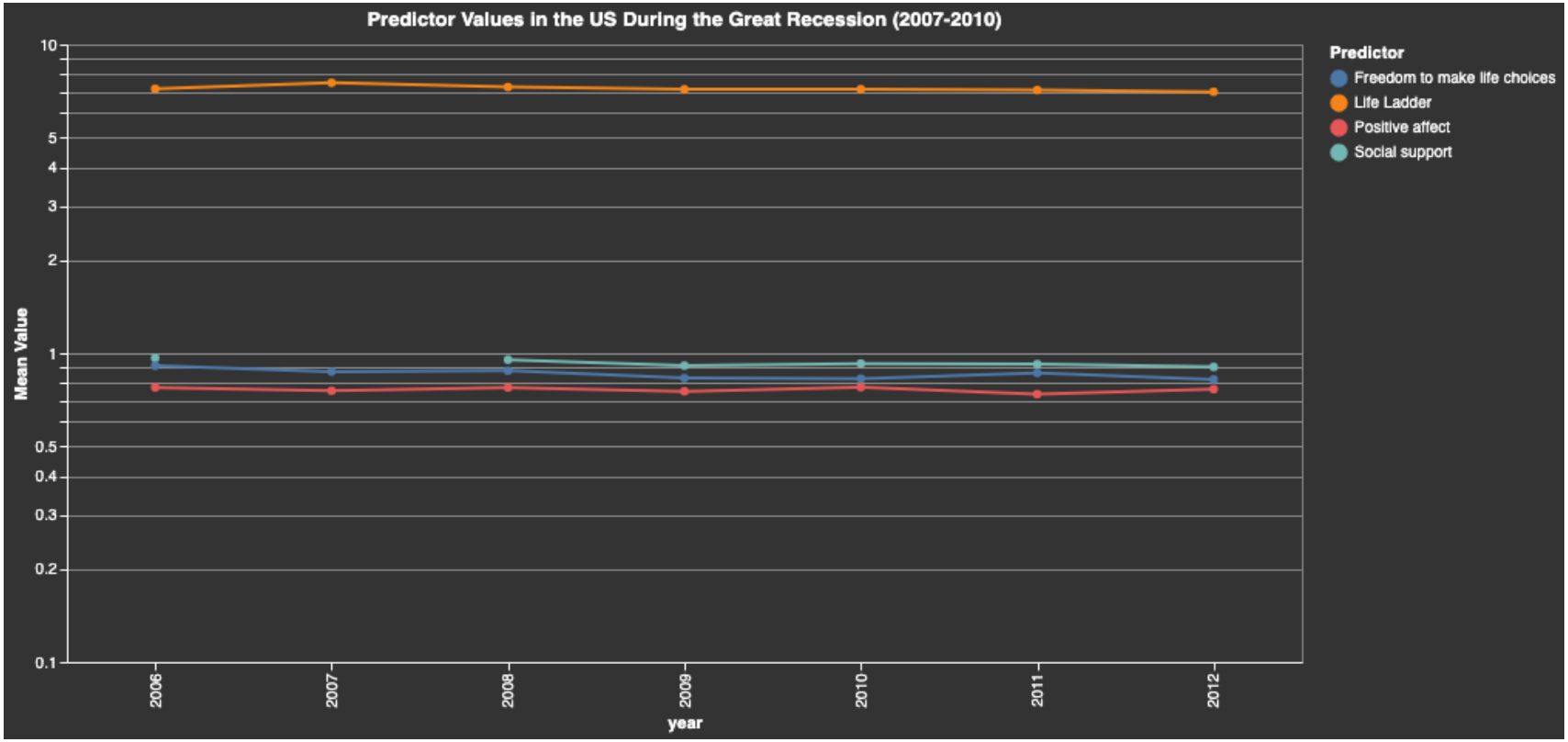
The happiest region had the least amount of countries while the least happy region had the greatest number of countries. One aspect that would be interesting to explore is how cultural homogenity affects happiness. It could be plausible that there is some racial/cultural tension that is not fully captured by this data set.

## Question 4: How did large scale events affect happiness?

### United States Great Recession

The first large scale event we decided to explore is the Great Recession in the United States

In [238…
```
display(mean_line_chart)
print('t-statistic:', t_stat)
print('p-value:', p_value)
```




```
t-statistic: 81.05059504862174
p-value: 5.293039104143237e-153
```

Just as we expected it seems that Life Ladder was on the constant decline starting in 2007, however we needed to make sure that this was a statistically significant decline so like before we calculated the p-value at a 0.05 significance level and we can confidently say that this decline was statistically significant.

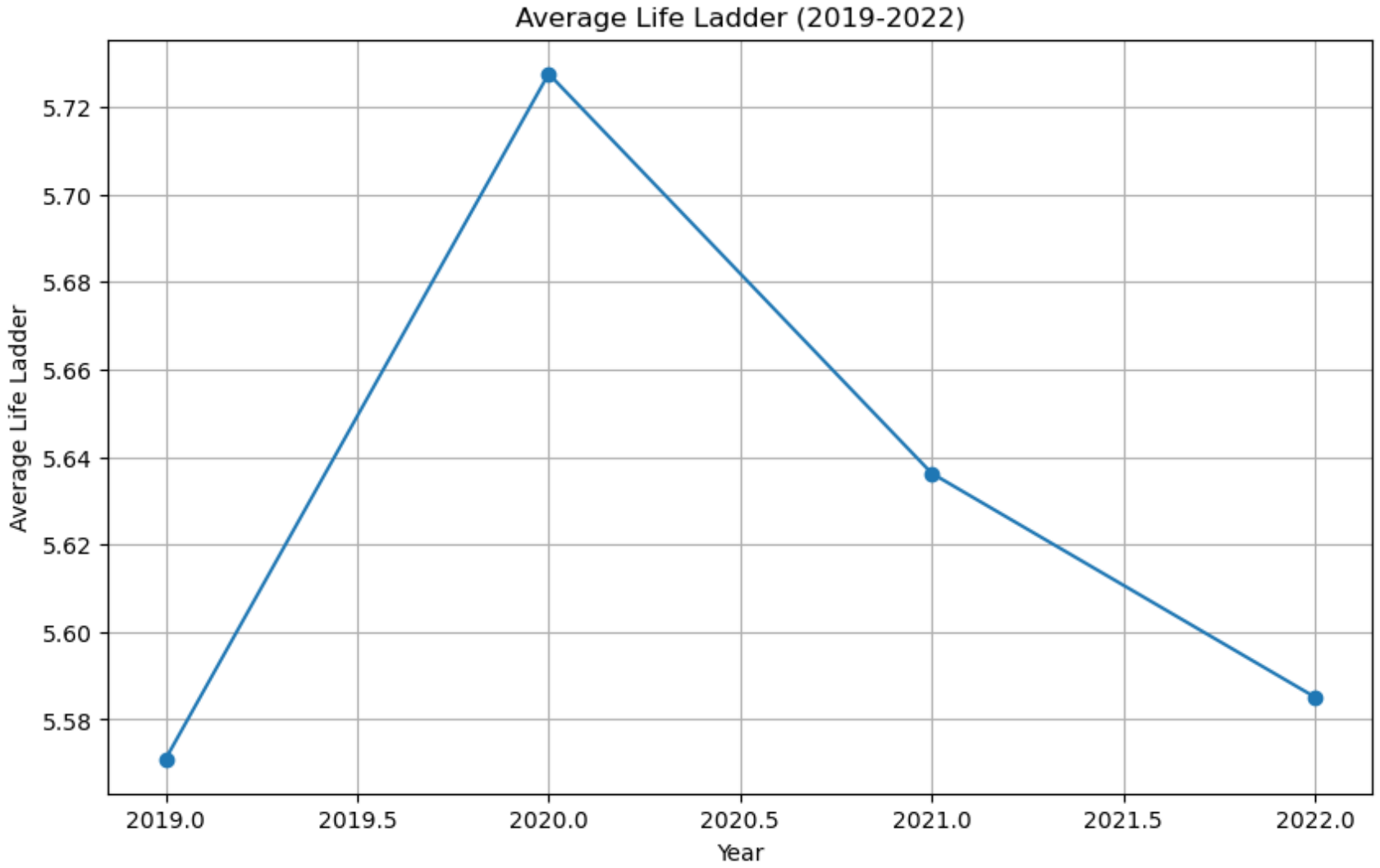
## 2020 COVID World Pandemic

The next event we were interested in analyzing is possibly the one event that comes to mind when people think of large scale event that affected happiness and that's the COVID Pandemic.

In [239…

```
average_positive_affect = whr[whr['year'].isin([2019, 2020, 2021, 2022])].groupby('year')['Life Ladder'].me

# Create a bar chart to visualize the average positive affect
# Create a line chart using matplotlib to visualize the average positive affect
```

```
plt.figure(figsize = (10, 6))
plt.plot(average_positive_affect['year'], average_positive_affect['Life Ladder'], marker = 'o')
plt.title('Average Life Ladder (2019-2022)')
plt.xlabel('Year')
plt.ylabel('Average Life Ladder')
plt.grid(True)
plt.show()
```




Based on the graphic above the trend seems clear that beginning in 2020 average life ladder values began a sharp decline.

In [240…
```
display(cases_with_whr.head(5))
```

| | Country | 2021 Covid | 2020 Life Ladder | 2021 Life Ladder | Change |
|---|---|---|---|---|---|
| **106** | Zambia | 228731 | 4.838 | 3.082 | -1.756 |
| **70** | Nepal | 567838 | 5.982 | 4.622 | -1.360 |
| **8** | Bangladesh | 1072029 | 5.280 | 4.123 | -1.157 |
| **74** | Nigeria | 154937 | 5.503 | 4.479 | -1.024 |
| **37** | Ghana | 91055 | 5.319 | 4.378 | -0.941 |

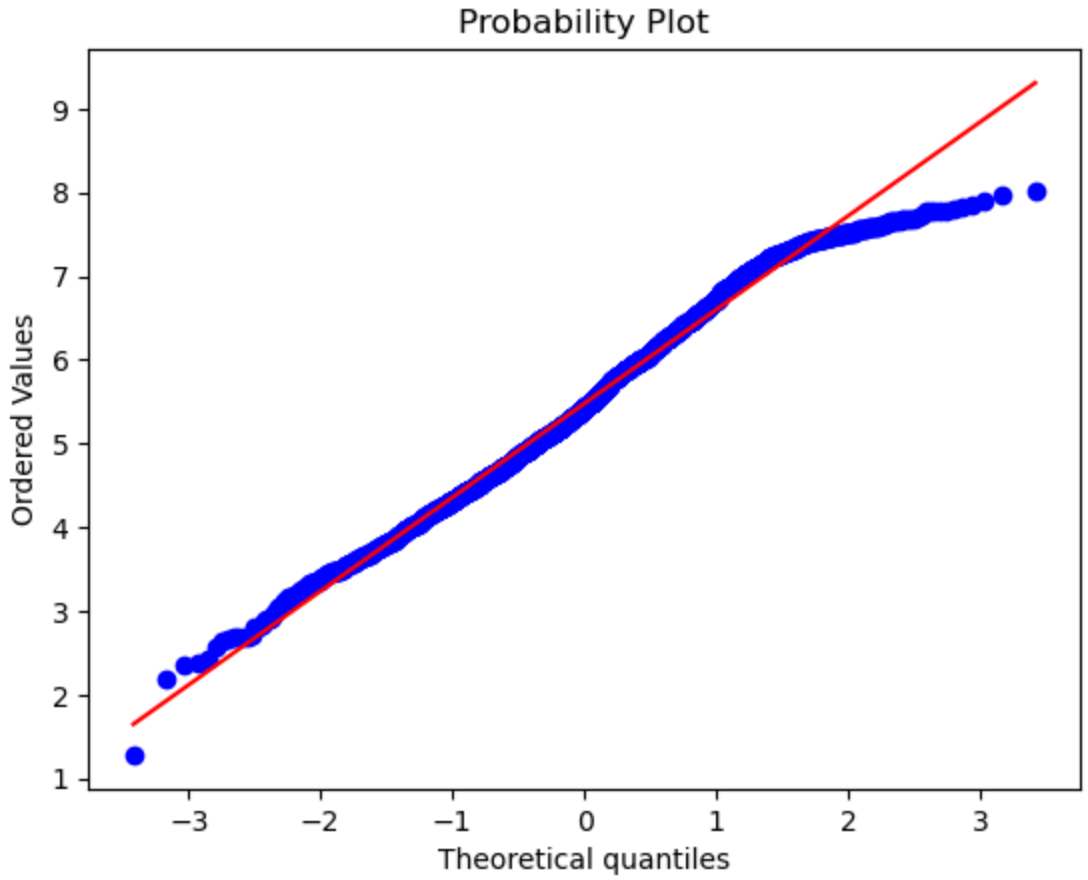
Continuing our analysis of regions, we constructed the table above will contains the country name, the number of covid cases in 2021, the life ladder value in 2020 and 2021, and the change in the life ladder between years. What you see above is the top 5 countries with the greatest decreases in life ladder values and in line with out previous analysis, all of the countries are in Asia and Africa which are the two regions with the lowest mean life ladder value.

## Question 5: Given a set of features, are we able to predict happiness using linear regression?

As the final step in our analysis we chose to fit a multiple linear regression model to the data and see how well we could predict "Life Ladder". The first step in our process was to check the assumptions for linearity and those are

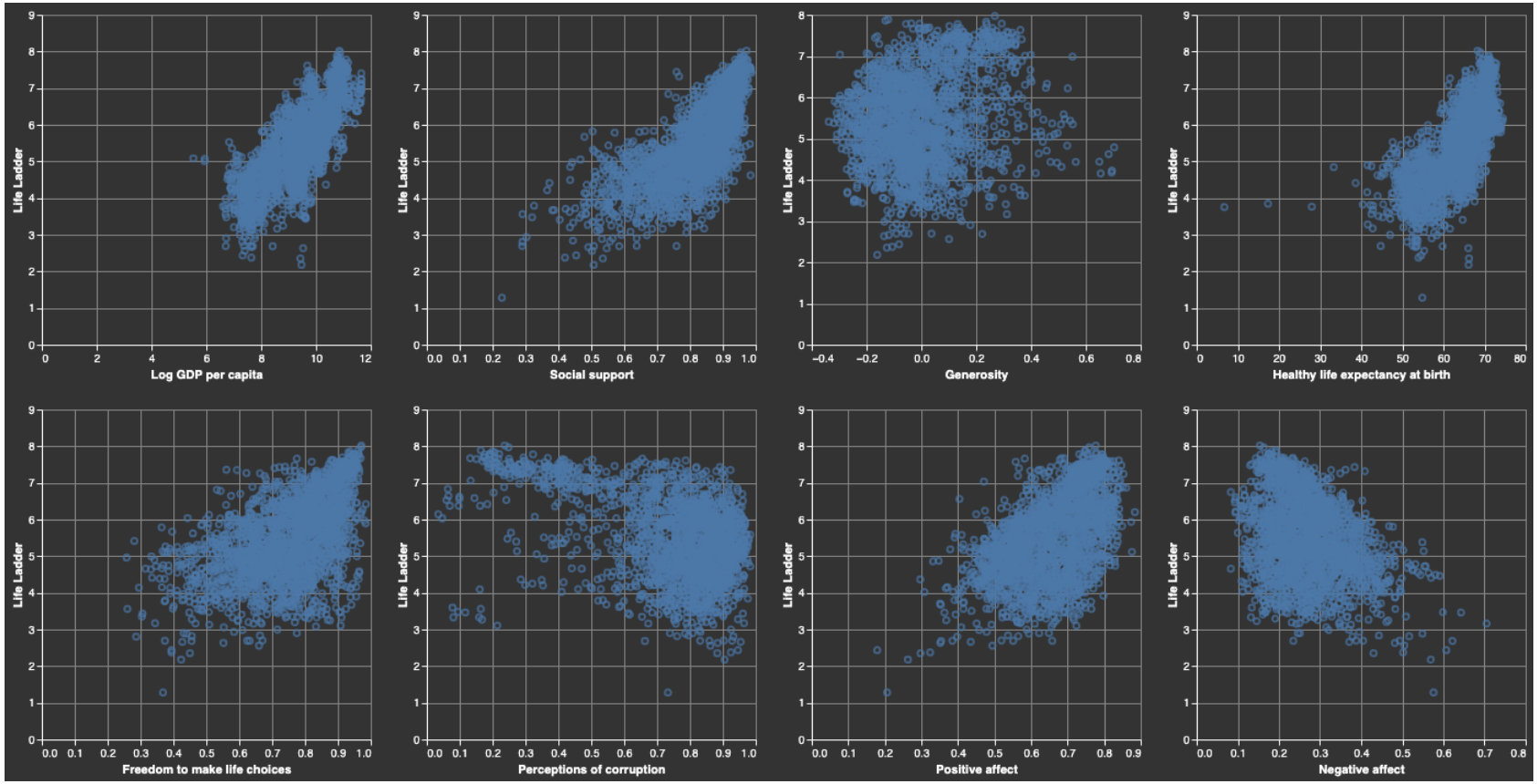
1. The response is normally distributed
2. The mean response is linear in the predictors
3. The observations are independent
4. The residuals are normally distributed with mean 0 and constant variance

In [241…
```
stats.probplot(whr["Life Ladder"], dist = "norm", plot = pylab)
pylab.show()
```




In [242…
```
life_ladder_scatter
```

Out[242…



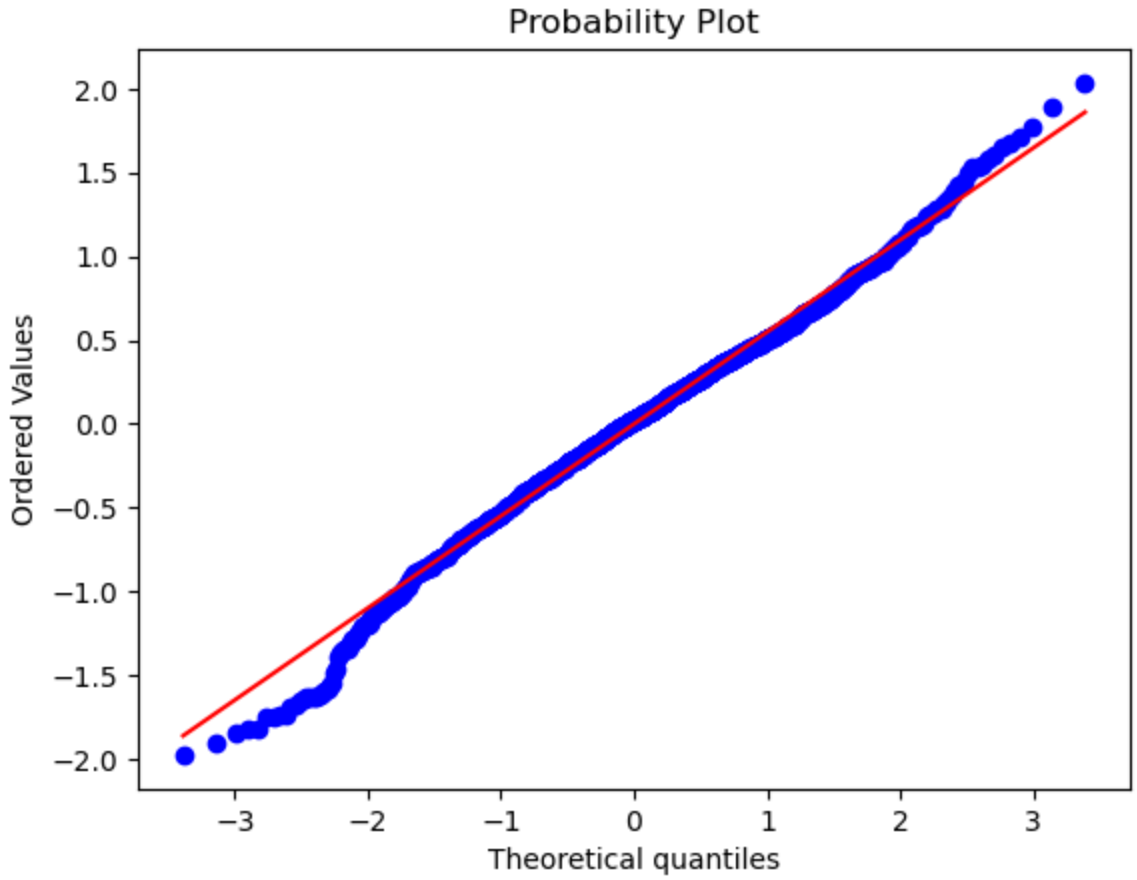

In [243…

```
print(vif_data)
```

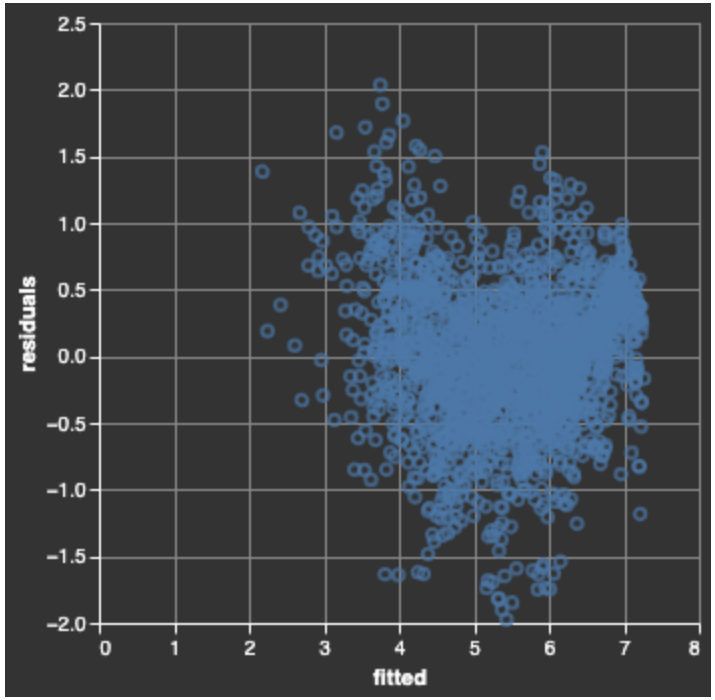
```
                        feature       VIF
0                Log GDP per capita  4.274933
1                    Social support  2.696033
2  Healthy life expectancy at birth  3.449044
3      Freedom to make life choices  1.924405
4         Perceptions of corruption  1.451232
5                   Positive affect  1.740433
6                   Negative affect  1.403463
```

In [244…

```
stats.probplot(res_fit_data["residuals"], dist="norm", plot=pylab)
pylab.show()
resid_plot
```


Probability Plot
Ordered Values
Theoretical quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−3
−2
−1
0
1
2
3

Out[244…




It seems from the plots that all of our assumptions are satisifed so we can fit the model. The full details can be found in the Code Appendix but below we have the coefficient estimates and our R-squared value.

In [245…

```
print(coef_tbl)
print('r-squared:', r_squared)
```

```
                                 estimate  standard error
const                           -3.683541        0.149475
Log GDP per capita               0.427353        0.021789
Social support                   1.506152        0.162006
Healthy life expectancy at birth 0.028011        0.003268
Freedom to make life choices     0.842994        0.115396
Positive affect                  2.476112        0.152632
Negative affect                 -0.307475        0.169844
error variance                   0.303916             NaN
r-squared: 0.7672797667494986
```

We have a final r-squared value of 0.76 which means that our model explains 76% of the variance in the data which is pretty good considering that we are trying to predict something as elusive as happiness.

## Conclusion

In conclusion, we find that both long term and short term happiness are affected by different variables. The prevailing factors affecting long term happiness seem to be more permanent in nature. Take for example "Log GDP per capita", something not easily changed and more of a national concern that is out of the control of the common folk. In contrast the prevailing factors affecting short term happiness seem to be things that are in control on an individual basis, such as a person's social circle and their feeling of freedom.

## Code Appendix

In [213…
```
# Reading in our datasets
whr_raw = pd.read_csv("data/whr-2023.csv")
regions = pd.read_csv("data/regions.csv")
covid = pd.read_csv("data/WHO-COVID-19-global-data.csv")
```

In [214…
```
# Merging our regions dataset with the whr data on "Country name"
regions.rename(columns = {'name':'Country name'}, inplace = True)
regions.drop(regions.columns.difference(['Country name','sub-region', 'region']), axis = 1, inplace = True
regions.head()

# Taking a look at the data
# It looks like it's already tidy so no further processing is necessary
whr = pd.merge(whr_raw, regions, how = 'left', on = 'Country name')
whr.head()
```

Out [214…

| | Country name | year | Life Ladder | Log GDP per capita | Social support | Healthy life expectancy at birth | Freedom to make life choices | Generosity | Perceptions of corruption | Positive affect | Negative affect | regior |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | 2008 | 3.724 | 7.350 | 0.451 | 50.5 | 0.718 | 0.168 | 0.882 | 0.414 | 0.258 | Asia |
| **1** | Afghanistan | 2009 | 4.402 | 7.509 | 0.552 | 50.8 | 0.679 | 0.191 | 0.850 | 0.481 | 0.237 | Asia |
| **2** | Afghanistan | 2010 | 4.758 | 7.614 | 0.539 | 51.1 | 0.600 | 0.121 | 0.707 | 0.517 | 0.275 | Asia |
| **3** | Afghanistan | 2011 | 3.832 | 7.581 | 0.521 | 51.4 | 0.496 | 0.164 | 0.731 | 0.480 | 0.267 | Asia |
| **4** | Afghanistan | 2012 | 3.783 | 7.661 | 0.521 | 51.7 | 0.531 | 0.238 | 0.776 | 0.614 | 0.268 | Asia |

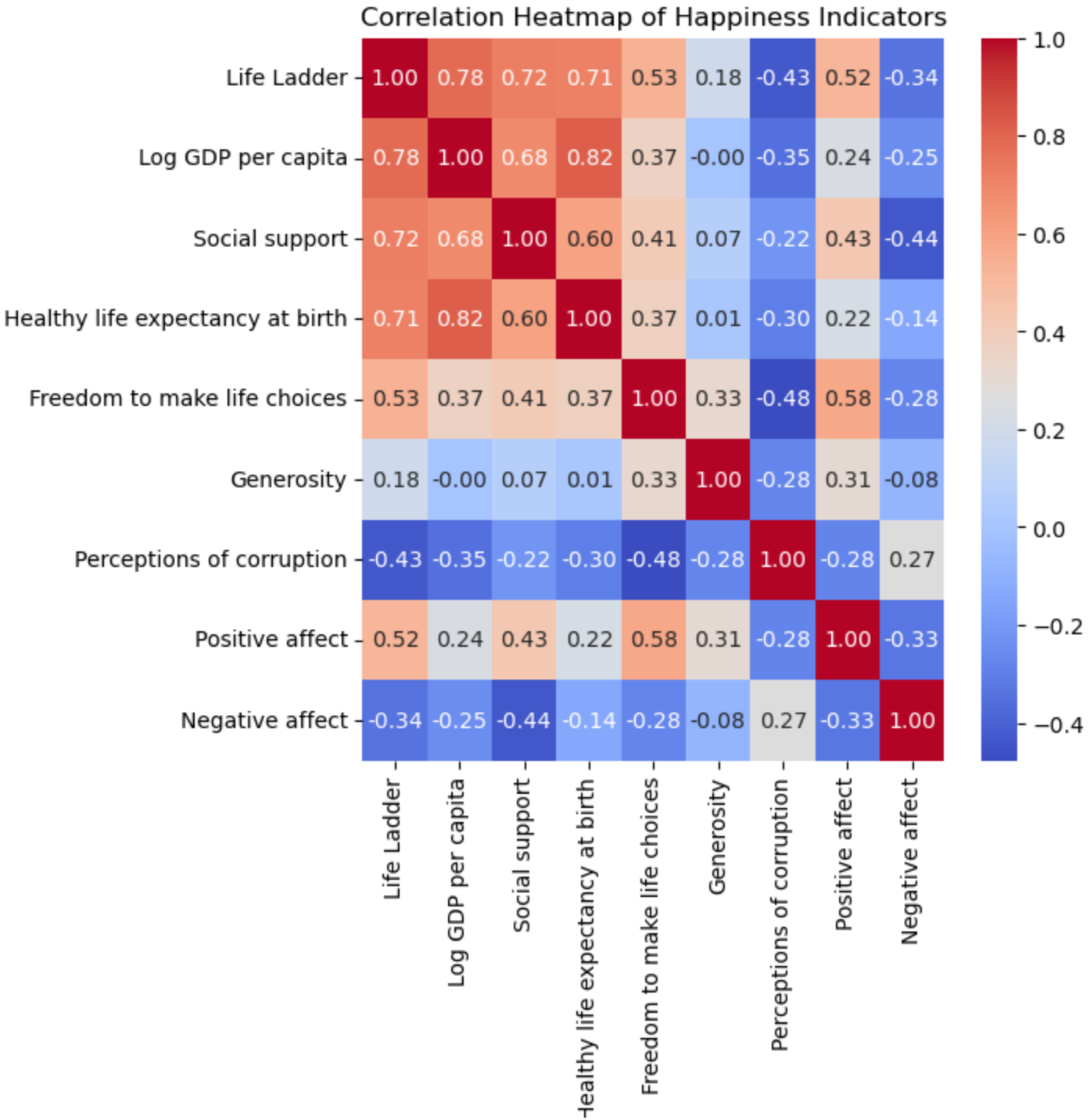
In [215…

```
# Checking the linear relationship between life ladder and predictors

fig_1 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Log GDP per capita")
)

fig_2 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Social support")
)

fig_3 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Generosity")
)

fig_4 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Healthy life expectancy at birth")
)
```

```
fig_5 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Freedom to make life choices")
)

fig_6 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Perceptions of corruption")
)

fig_7 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Positive affect")
)

fig_8 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Negative affect")
)

upper = fig_1 | fig_2 | fig_3 | fig_4
lower = fig_5 | fig_6 | fig_7 | fig_8

life_ladder_scatter = alt.vconcat(upper, lower)
```

In [216…

```
# Correlation matrix
corr = whr.select_dtypes(include = ['float64']).corr()

# Plot the heatmap
plt.figure(figsize = (6, 6))
sn.heatmap(corr, annot = True, fmt = '.2f', cmap = 'coolwarm')
plt.title('Correlation Heatmap of Happiness Indicators')
plt.show()
```


Correlation Heatmap of Happiness Indicators
Life Ladder
Log GDP per capita
Social support
Healthy life expectancy at birth
Freedom to make life choices
Generosity
Perceptions of corruption
Positive affect
Negative affect
1.00 0.78 0.72 0.71 0.53 0.18 -0.43 0.52 -0.34
0.78 1.00 0.68 0.82 0.37 -0.00 -0.35 0.24 -0.25
0.72 0.68 1.00 0.60 0.41 0.07 -0.22 0.43 -0.44
0.71 0.82 0.60 1.00 0.37 0.01 -0.30 0.22 -0.14
0.53 0.37 0.41 0.37 1.00 0.33 -0.48 0.58 -0.28
0.18 -0.00 0.07 0.01 0.33 1.00 -0.28 0.31 -0.08
-0.43 -0.35 -0.22 -0.30 -0.48 -0.28 1.00 -0.28 0.27
0.52 0.24 0.43 0.22 0.58 0.31 -0.28 1.00 -0.33
-0.34 -0.25 -0.44 -0.14 -0.28 -0.08 0.27 -0.33 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

In [217…
```
# Further exploring the strongest relationships with Life Ladder
data_2009_2012 = whr[whr['year'].isin([2009, 2012])]

life_gdp = alt.Chart(data_2009_2012).mark_circle(opacity = 0.5).encode(
    x = alt.X('Log GDP per capita:Q'),
    y = alt.Y('Life Ladder', title = 'Life Ladder', scale = alt.Scale(zero = False)),
    color = 'region',
    size = 'Generosity'
).properties(
    width = 600,
    height = 400
).configure_axis(
    grid = False
)
```

In [218…
```
# Calculate the heatmap for positive affect

corr_matrix = whr.select_dtypes(include = ['float64']).corr()

# Create a dataframe for a heatmap
corr_df = corr_matrix.reset_index().melt('index')

# Create the heatmap
heatmap = alt.Chart(corr_df).mark_rect().encode(
    x = 'index:O',
    y = 'variable:O',
    color = 'value:Q'
).properties(
    title = 'Correlation Heatmap',
    width = 400,
    height = 50
)

# Filter to only show correlations for 'Life Ladder'
positive_affect_heatmap = heatmap.transform_filter(
    alt.datum.variable == 'Positive affect'
)
```

In [219…
```
# Calculate the mean life ladder by year and subregion
mean_life_ladder = whr.groupby(['year', 'region'])['Life Ladder'].mean().reset_index()
```

```
# Creating plot
life_ladder_region = alt.Chart(mean_life_ladder).mark_line().encode(
    x = alt.X('year:O', scale = alt.Scale(zero = False)),
    y = alt.Y('Life Ladder:Q'),
    color = 'region:N',
    tooltip = ['year', 'region', 'Life Ladder']
).properties(
    title = 'Mean Life Ladder by Year and Region'
)
```

In [220…

```
# Define the groups
regions_group_1 = ['Oceania']
regions_group_2 = ['Africa']

# Filter the dataset into two groups
whr_group_1 = whr[whr['region'].isin(regions_group_1)]
whr_group_2 = whr[whr['region'].isin(regions_group_2)]

# Calculate the mean of the required columns for each group
mean_values_group_1 = whr_group_1.mean(numeric_only = True).reset_index()
mean_values_group_1['Region'] = 'Oceania'

mean_values_group_2 = whr_group_2.mean(numeric_only = True).reset_index()
mean_values_group_2['Region'] = 'Africa'

# Combine the groups
combined_means = pd.concat([mean_values_group_1, mean_values_group_2])

combined_means = combined_means.rename(columns = {'index':'Statistic', 0:'Value'})
combined_means = combined_means.reset_index()

relevant_indicators = ['Social support',
                       'Freedom to make life choices', 'Generosity',
                       'Perceptions of corruption', 'Positive affect','Negative affect']

combined_means_mod1 = combined_means[combined_means['Statistic'].isin(relevant_indicators)]

# Create the Altair chart
most_least_happy = alt.Chart(combined_means_mod1).mark_bar().encode(
    x = alt.X('Statistic:N', axis = None),
    y = alt.Y('Value:Q'),
```

```
    color = 'Statistic:N',
    column = 'Region:N',
    tooltip = ['Region', 'Statistic', 'Value']
).properties(width = 200)
```

In [221…
```
# Counting the number of countries which each region and displaying the count
countries_per_region = whr.groupby('region')['Country name'].nunique()
```

In [222…
```
# Filter the data for the Americas and the years 2007-2010
americas_data = whr[(whr['Country name'] == 'United States') &
                                (whr['year'] >= 2005) &
                                (whr['year'] <= 2012)]

# Calculate the mean of the predictors for each year
americas_mean_data = americas_data.groupby('year').mean(numeric_only = True).reset_index()

# Melt the dataframe to have a long format for Altair's lineplot
americas_mean_melted = americas_mean_data.melt(id_vars = ['year'],
                                               value_vars = ['Life Ladder', 'Social support',
                                                             'Freedom to make life choices',
                                                             'Positive affect'],
                                               var_name = 'Predictor', value_name='Mean Value')

# Create a lineplot for the mean value of each predictor through time using Altair
mean_line_chart = alt.Chart(americas_mean_melted).mark_line(point = True).encode(
    x = 'year:O',
    y = alt.Y('Mean Value:Q', scale=alt.Scale(type='log')),
    color = 'Predictor:N',
    tooltip = ['year', 'Predictor', 'Mean Value']
).properties(
    width = 800,
    height = 400,
    title = 'Predictor Values in the US During the Great Recession (2007-2010)'
).interactive()
```

In [223…
```
# Extract the 'Life Ladder' data for the United States and the rest of the world for the years 2008 and 20
us_life_ladder = whr[(whr['Country name'] == 'United States') &
                                     (whr['year'].isin([2008, 2009])) &
                                     (whr['Life Ladder'].notna())]['Life Ladder']

world_life_ladder = whr[(whr['Country name'] != 'United States') &
```

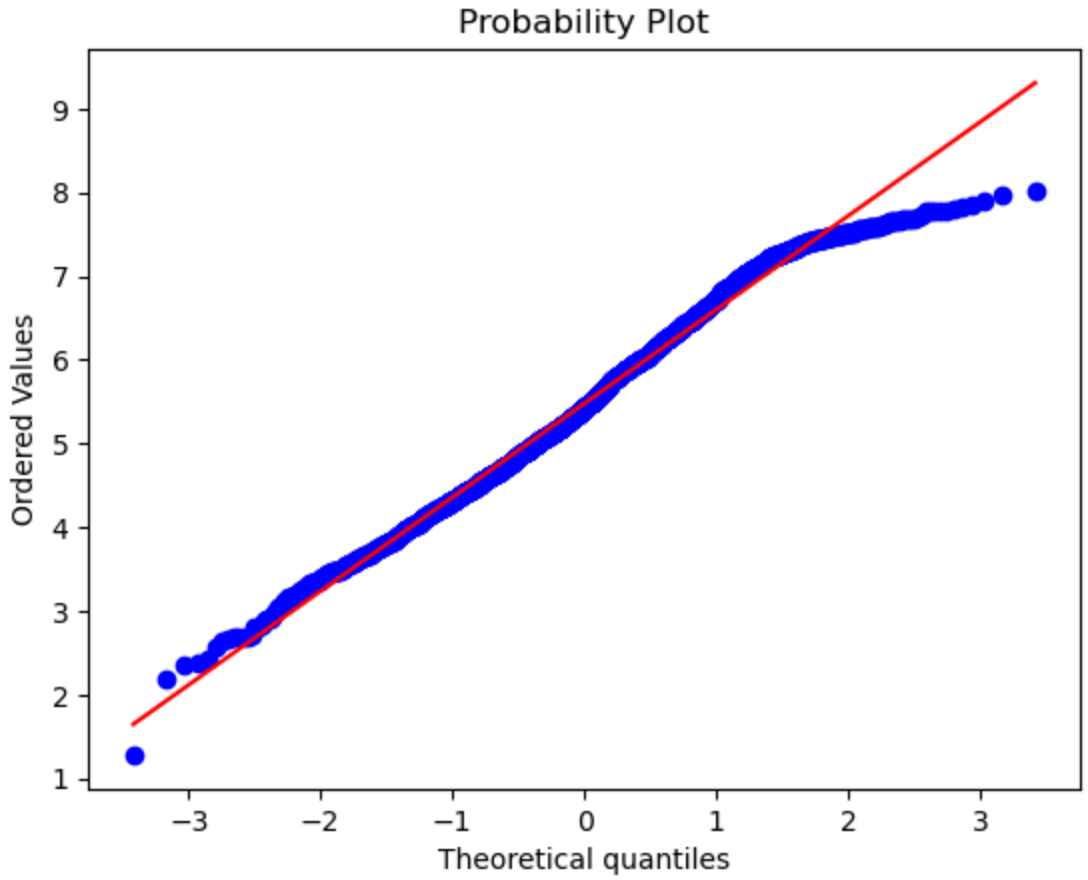
```
                                        (whr['year'].isin([2008, 2009])) &
                                        (whr['Life Ladder'].notna())]['Life Ladder']

# Perform a t-test to see if the decline is statistically significant
# Null hypothesis: There is no difference in life ladder between the United States and the rest of the wor
# Alternative hypothesis: There is a difference in life ladder between the United States and the rest of t

# Perform a two-sample t-test
t_stat, p_value = stats.ttest_ind(us_life_ladder, world_life_ladder, equal_var = False)
```

In [224…

```
# Filter the dataset for the year 2021
covid_2021 = covid[covid['Date_reported'].str.startswith('2021')]

# Calculate the total number of cases for each country in 2021
total_cases_2021 = covid_2021.groupby('Country')['New_cases'].sum().reset_index()

# Calculate the change in life ladder for each country between 2019 and 2020
# Filter the dataset for the years 2019 and 2020
years_filter = whr['year'].isin([2021, 2020])

# Pivot the table to have countries as rows and years as columns for positive affect
positive_affect_pivot = whr[years_filter].pivot(index = 'Country name', columns = 'year', values = 'Life La

# Calculate the change between the years
positive_affect_pivot['Change'] = positive_affect_pivot[2021] - positive_affect_pivot[2020]

# Sort the countries by the largest drop in positive affect
largest_drops = positive_affect_pivot.sort_values('Change')

# Display the countries with the largest drops
cases_with_whr = total_cases_2021.merge(largest_drops, left_on = 'Country', right_on = 'Country name',
                                        how = 'inner').rename(columns = {'New_cases':'2021 Covid',
                                                                         2020:'2020 Life Ladder',
                                                                         2021:'2021 Life Ladder'}).sort_va
```

In [225…

```
# Checking that the response is normal
stats.probplot(whr["Life Ladder"], dist = "norm", plot = pylab)
pylab.show()
```




In [226…

```
# Checking the linear relationship between life ladder and predictors
# Predictors were picked using the correlation matrix

fig_1 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Log GDP per capita")
)

fig_2 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Social support")
)

fig_3 = alt.Chart(whr).mark_point().encode(
```

```
    y = alt.X("Life Ladder"),
    x = alt.X("Generosity")
)

fig_4 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Healthy life expectancy at birth")
)

fig_5 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Freedom to make life choices")
)

fig_6 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Perceptions of corruption")
)

fig_7 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Positive affect")
)

fig_8 = alt.Chart(whr).mark_point().encode(
    y = alt.X("Life Ladder"),
    x = alt.X("Negative affect")
)

upper = fig_1 | fig_2 | fig_3 | fig_4
lower = fig_5 | fig_6 | fig_7 | fig_8

life_ladder_scatter = alt.vconcat(upper, lower)

# Checking predictors for colinearity
X = whr[['Log GDP per capita', 'Social support',
         'Healthy life expectancy at birth', 'Freedom to make life choices',
         "Perceptions of corruption", 'Positive affect', 'Negative affect']].dropna()

# VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns
```

```
# calculating VIF for each feature
vif_data["VIF"] = [variance_inflation_factor(X.values, i)
                          for i in range(len(X.columns))]

print(vif_data)

# Found extreme dependence among predictors so centering and scaling them to remove dependence
X = whr[['Log GDP per capita', 'Social support',
         'Healthy life expectancy at birth', 'Freedom to make life choices',
         "Perceptions of corruption", 'Positive affect', 'Negative affect']].dropna().apply(lambda x: x - x

# VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns

# Calculating VIF for each feature
vif_data["VIF"] = [variance_inflation_factor(X.values, i)
                          for i in range(len(X.columns))]

print(vif_data)
```

```
                        feature         VIF
0               Log GDP per capita  271.014563
1                   Social support  118.827235
2  Healthy life expectancy at birth  272.262939
3      Freedom to make life choices   54.156019
4         Perceptions of corruption   17.219028
5                   Positive affect   59.880904
6                   Negative affect   13.959606
                        feature       VIF
0               Log GDP per capita  4.274933
1                   Social support  2.696033
2  Healthy life expectancy at birth  3.449044
3      Freedom to make life choices  1.924405
4         Perceptions of corruption  1.451232
5                   Positive affect  1.740433
6                   Negative affect  1.403463
```

In [227…

```
# Creating regression data
regression_data = whr.copy().dropna()
```

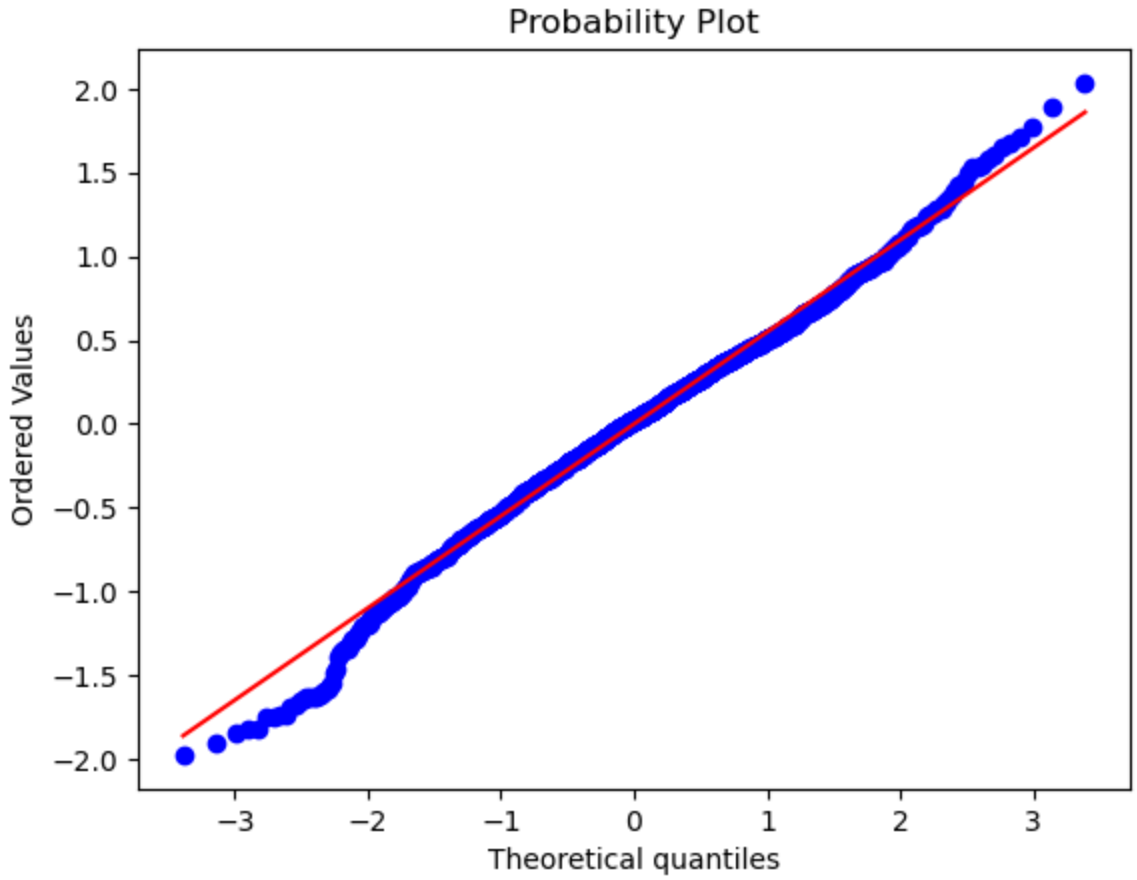
```
x_vars = regression_data[['Log GDP per capita', 'Social support',
                          'Healthy life expectancy at birth', 'Freedom to make life choices',
                          'Positive affect', 'Negative affect']]

x = sm.tools.add_constant(x_vars)
y = regression_data["Life Ladder"]

# Fitting MLR Model
mlr = sm.OLS(endog = y, exog = x)
rslt = mlr.fit()

# retreive estimates and std errors
coef_tbl = pd.DataFrame({
    'estimate': rslt.params,
    'standard error': np.sqrt(rslt.cov_params().values.diagonal())
}, index = x.columns.values)

coef_tbl.loc['error variance', 'estimate'] = rslt.scale
```

In [228…

```
fitted_values = rslt.fittedvalues
residuals = rslt.resid
r_squared = rslt.rsquared
```

In [229…

```
# Checking that residuals are normally distributed with mean 0 and constant variance
data = {"fitted": fitted_values,
                "residuals": residuals}

res_fit_data = pd.concat(data, axis = 1)

resid_plot = alt.Chart(res_fit_data).mark_point().encode(
         x = alt.X("fitted:Q"),
         y = alt.Y("residuals:Q")
)

stats.probplot(res_fit_data["residuals"], dist="norm", plot=pylab)
pylab.show()
```


Probability Plot
Ordered Values
Theoretical quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−3
−2
−1
0
1
2
3