Final Project

Happiness ~ Life Expectancy Relationship

1/28/2021

Jeff Schmidt -- Database & Data Exploration
Duc Luu -- Machine Learning & Analysis Result
JP Enciso-Siller -- Dashboard & Github Administration

Outline

- Selected Topic & Reason Selected (Jeff)
- Description of the Source Data (Jeff)
- Questions We Hope to Answer With the Data (Duc)
- Data Exploration (JP)
- Analysis & Machine Learning Phase (Duc)
- Dashboard Review (JP)
- Github Administration (JP)

Selected Topic & Reason Selected

Topic: The effect of various factors on happiness.

What factors affect happiness?

Does life expectancy impact happiness?

What factors do we think affect happiness but actually don't?

Reason: Identify correlation between independent data sets.

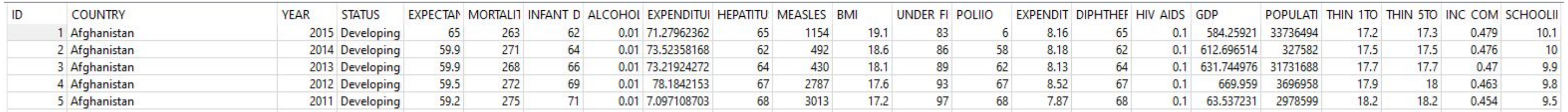
Source Data:

- https://www.kaggle.com/kumarajarshi/life-expectancy-who
- https://www.kaggle.com/unsdsn/world-happiness

Data sets are from differing sources covering separate time periods.

However, seen as an opportunity to identify same relation existing despite source.

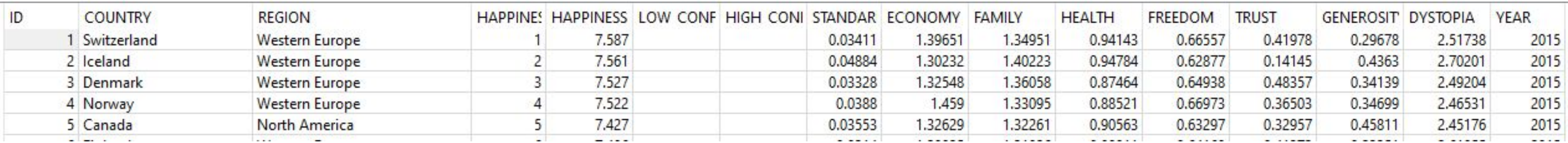
Description of the Source Data



ID	COUNTRY	YEAR	STATUS	EXPECTAN	MORTALIT	INFANT D	ALCOHOL	EXPENDITU	HEPATITU	MEASLES	BMI	UNDER FI	POLIIO	EXPENDIT	DIPHTHEF	HIV AIDS	GDP	POPULATI	THIN 1TO	THIN 5TO	INC COM	SCHOOLI
1	Afghanistan	2015	Developing	65	263	62	0.01	71.27962362	65	1154	19.1	83	6	8.16	65	0.1	584.25921	33736494	17.2	17.3	0.479	10.1
2	Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358168	62	492	18.6	86	58	8.18	62	0.1	612.696514	327582	17.5	17.5	0.476	10
3	Afghanistan	2013	Developing	59.9	268	66	0.01	73.21924272	64	430	18.1	89	62	8.13	64	0.1	631.744976	31731688	17.7	17.7	0.47	9.9
4	Afghanistan	2012	Developing	59.5	272	69	0.01	78.1842153	67	2787	17.6	93	67	8.52	67	0.1	669.959	3696958	17.9	18	0.463	9.8
5	Afghanistan	2011	Developing	59.2	275	71	0.01	7.097108703	68	3013	17.2	97	68	7.87	68	0.1	63.537231	2978599	18.2	18.2	0.454	9.5

Life Expectancy

- Column count = 22 (exclude ID); Row count = 2938
- Key columns: EXPECTANCY, STATUS, COUNTRY
- Timeframe = 2000 ~ 2015
- Source = World Health Organization & United Nations

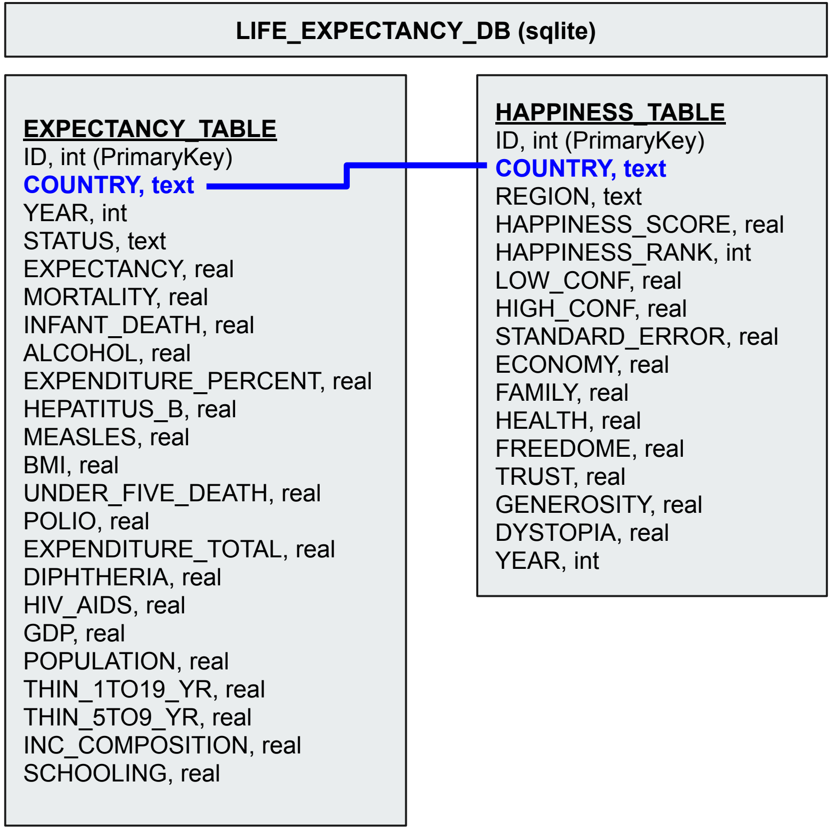


ID	COUNTRY	REGION	HAPPINES	HAPPINESS	LOW CONF	HIGH CONI	STANDAR	ECONOMY	FAMILY	HEALTH	FREEDOM	TRUST	GENEROSIT	DYSTOPIA	YEAR
1	Switzerland	Western Europe	1	7.587			0.03411	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2.51738	2015
2	Iceland	Western Europe	2	7.561			0.04884	1.30232	1.40223	0.94784	0.62877	0.14145	0.4363	2.70201	2015
3	Denmark	Western Europe	3	7.527			0.03328	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	2.49204	2015
4	Norway	Western Europe	4	7.522			0.0388	1.459	1.33095	0.88521	0.66973	0.36503	0.34699	2.46531	2015
5	Canada	North America	5	7.427			0.03553	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	2.45176	2015

World Happiness

- Column count = 15 (exclude ID); Row count = 782
- Key columns: Happiness Rank, Happiness Score, Health(Life Expectancy), Country
- Timeframe = 2015 ~ 2019
- Source = Gallup World Poll

Description of the Source Data (cont)



LIFE_EXPECTANCY_DB (sqlite)

EXPECTANCY_TABLE
ID, int (PrimaryKey)
COUNTRY, text
YEAR, int
STATUS, text
EXPECTANCY, real
MORTALITY, real
INFANT_DEATH, real
ALCOHOL, real
EXPENDITURE_PERCENT, real
HEPATITUS_B, real
MEASLES, real
BMI, real
UNDER_FIVE_DEATH, real
POLIO, real
EXPENDITURE_TOTAL, real
DIPHTHERIA, real
HIV_AIDS, real
GDP, real
POPULATION, real
THIN_1TO19_YR, real
THIN_5TO9_YR, real
INC_COMPOSITION, real
SCHOOLING, real

HAPPINESS_TABLE
ID, int (PrimaryKey)
COUNTRY, text
REGION, text
HAPPINESS_SCORE, real
HAPPINESS_RANK, int
LOW_CONF, real
HIGH_CONF, real
STANDARD_ERROR, real
ECONOMY, real
FAMILY, real
HEALTH, real
FREEDOME, real
TRUST, real
GENEROSITY, real
DYSTOPIA, real
YEAR, int

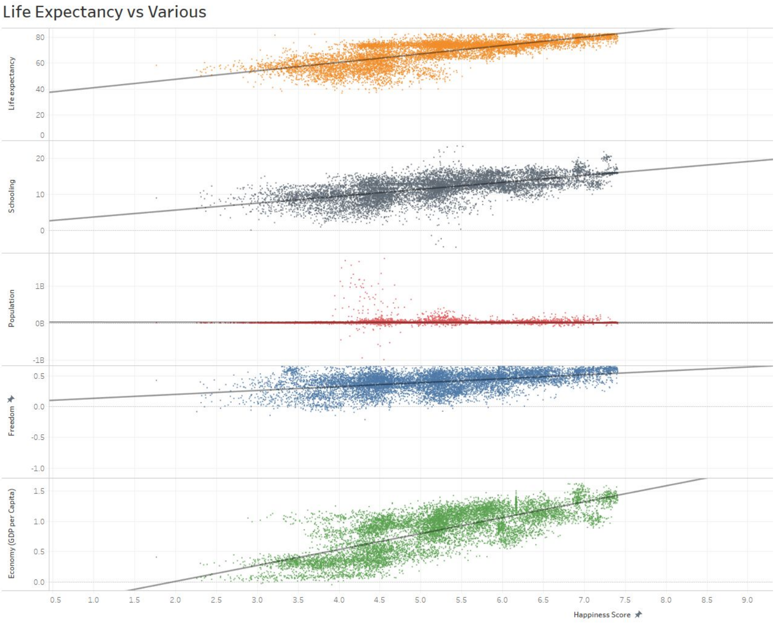
- Issues with raw .csv files
  - No distinct key column
  - Inconsistent naming convention across files
  - Poor SQL naming convention within file
- Two tables formed under single schema.
- View created using simple inner join based on country name & common year.
  - INNER JOIN HAPPINESS_TABLE H ON UPPER(H.COUNTRY = UPPER(E.COUNTRY) WHERE E.YEAR = ‘2015’ and H.YEAR = 2015
- Data saved to .db file format for load via SQLALCHEMY into Pandas Dataframe
- Read all requirements & problem statements before hand--i.e. Take time with data gathering
  - Data selection was initially influenced based on what was of interest, capability to produce a result, and level of interest.
  - However, project data requirements expanded during 2nd deliverable (2 tables & inclusion of a JOIN).
    - This requirement would have influenced data selection during 1st week deliverable.

Questions Hope to Answer with the Data

- Which health factors have the biggest impact of happiness?
- Which economic factors have a biggest impact on happiness?
- What factors would we expect to affect happiness but actually does not?
- Which country had the highest/lowest predicted happiness score?

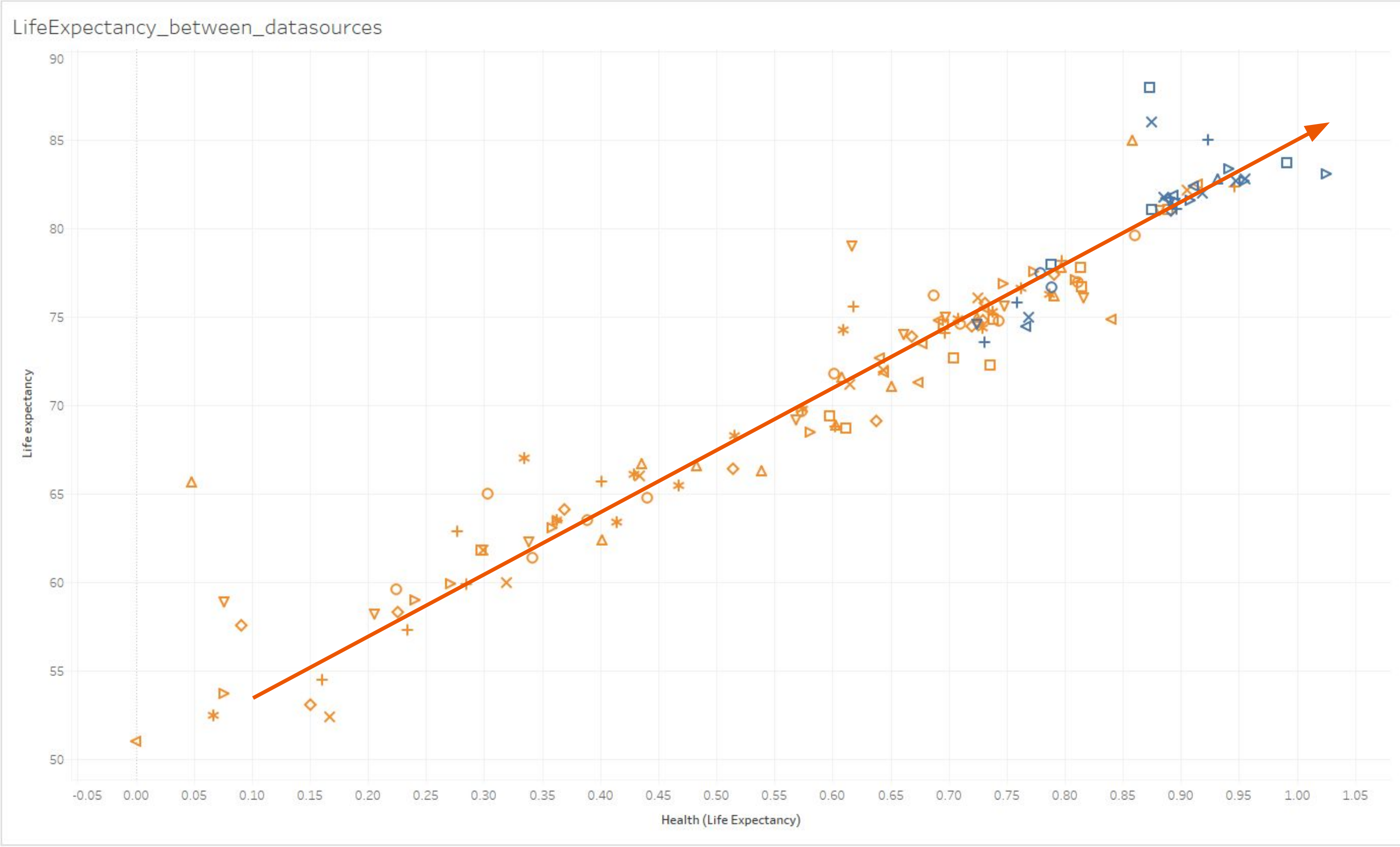
Data Exploration

- Use of Tableau to plot raw data for variable columns against “Happiness Score” to find relationships
- Plotted to gain a visual understanding prior to application of Machine Learning.
    - Early identification of potential correlation of happiness
    - Used as a tool to identify points of merging data
    - Identify what data needs cleaning and processing (ex null values, missing data, possible mistakes)



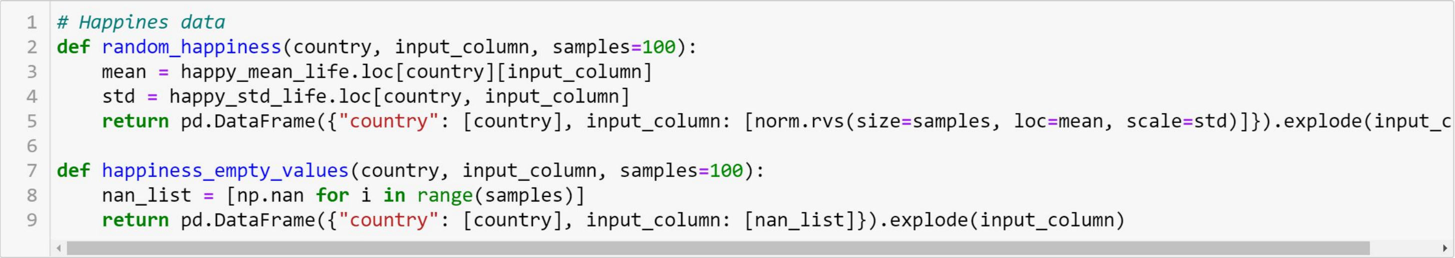
Data Exploration (cont.)

- Joined data between Life Expectancy & Happiness could only be done for year 2015.
- Happiness data utilizes “score” for Life Expectancy whereas Life Expectancy data uses actual age.
  - Visualization used to confirm linear relation between score & age.



Analysis & Machine Learning Phase

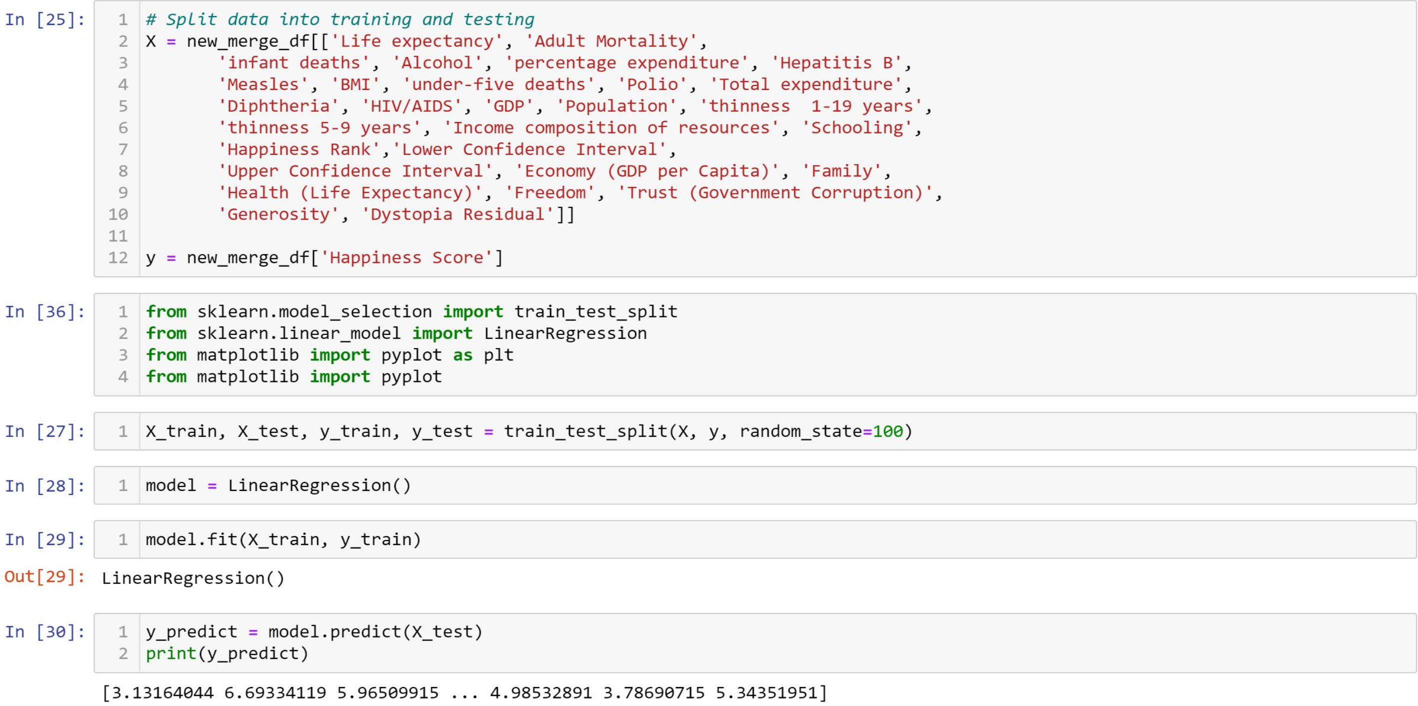
- Data Issues
  - Filling Blanks
  - Explode Data
- Which Model Used?



```python
# Happines data
def random_happiness(country, input_column, samples=100):
    mean = happy_mean_life.loc[country][input_column]
    std = happy_std_life.loc[country, input_column]
    return pd.DataFrame({"country": [country], input_column: [norm.rvs(size=samples, loc=mean, scale=std)]}).explode(input_c

def happiness_empty_values(country, input_column, samples=100):
    nan_list = [np.nan for i in range(samples)]
    return pd.DataFrame({"country": [country], input_column: [nan_list]}).explode(input_column)
```

Analysis & Machine Learning Phase



In [25]:
```
# Split data into training and testing
X = new_merge_df[['Life expectancy', 'Adult Mortality',
       'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatitis B',
       'Measles', 'BMI', 'under-five deaths', 'Polio', 'Total expenditure',
       'Diphtheria', 'HIV/AIDS', 'GDP', 'Population', 'thinness  1-19 years',
       'thinness 5-9 years', 'Income composition of resources', 'Schooling',
       'Happiness Rank','Lower Confidence Interval',
       'Upper Confidence Interval', 'Economy (GDP per Capita)', 'Family',
       'Health (Life Expectancy)', 'Freedom', 'Trust (Government Corruption)',
       'Generosity', 'Dystopia Residual']]

y = new_merge_df['Happiness Score']
```

In [36]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from matplotlib import pyplot as plt
from matplotlib import pyplot
```

In [27]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=100)
```

In [28]:
```
model = LinearRegression()
```

In [29]:
```
model.fit(X_train, y_train)
```

Out[29]:
```
LinearRegression()
```

In [30]:
```
y_predict = model.predict(X_test)
print(y_predict)
```

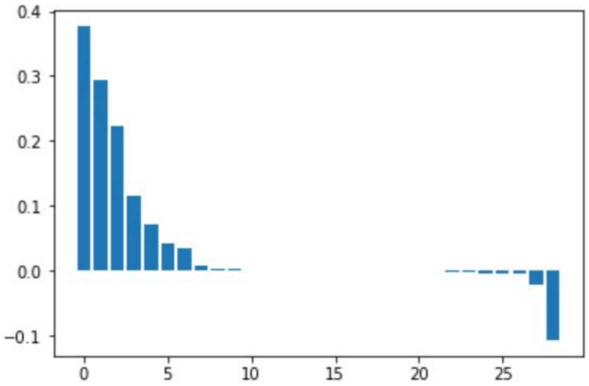
```
[3.13164044 6.69334119 5.96509915 ... 4.98532891 3.78690715 5.34351951]
```

Analysis & Machine Learning Phase (cont.)

- Economic/Health Factors
- Country with the lowest predicted happiness score:
    - Burundi: 3.0665
- Country with the highest predicted happiness score:
    - Netherlands: 7.4531
- Future Analysis

Analysis & Machine Learning Phase (cont.)

Results of Analysis



Feature Importance Graph

- Highest Health Factors:
  - Alcohol: 0.00282
  - Life Expectancy: 0.00270
  - Family: 0.03388
- Highest Economic Factors:
  - Freedom: 0.22304
  - Generosity: 0.11409
  - Economy (GDP per Capita): 0.07213
- Lowest Health Factors:
  - Health (life expectancy): -0.02122
  - Thinness 5-9 years: -0.00459
  - HIV/AIDS: -0.00431
- Lowest Economic Factors:
  - Trust (Government Corruption): -0.10733
  - Schooling: -0.00167
  - GDP: 0.000

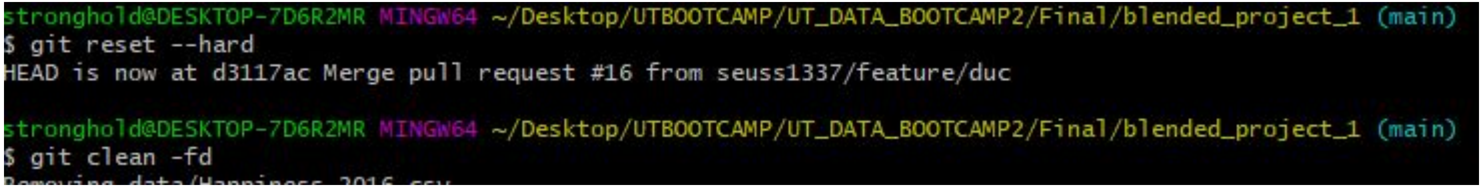
Dashboard Review

Dashboard

- Originally started with Flask transitioned to Tableau
- Shows model predictions vs actual test data
- Gives a visual representation of what the underlying data looks like
- https://public.tableau.com/profile/jp4411#!/vizhome/SemiFinalDashboard/Dashboard1

Github Administration

- Project Management Role
  - Being the go to git resource
  - Assisting teammates with github errors
  - Coordinating with team members about their piece of the project
- Merge Conflict Resolution
- Lessons Learned
  - Communication is essential
  - All team members getting in the habit of working with branches
  - Establishing a folder structure early on
  - GOOGLE GOOGLE GOOGLE



```
stronghold@DESKTOP-7D6R2MR MINGW64 ~/Desktop/UTBOOTCAMP/UT_DATA_BOOTCAMP2/Final/blended_project_1 (main)
$ git reset --hard
HEAD is now at d3117ac Merge pull request #16 from seuss1337/feature/duc

stronghold@DESKTOP-7D6R2MR MINGW64 ~/Desktop/UTBOOTCAMP/UT_DATA_BOOTCAMP2/Final/blended_project_1 (main)
$ git clean -fd
Removing data/Happiness_2016.csv
```