ETL PROJECT

EXTRACTION, TRANSFORMATION AND LOAD

Manuela Hoyos
Sura Baghirova
Gilbert Fevry

SQL SCHEMA AND QUERY

A relational database called 'ETL_db' was created in PostgresSQL pgAdmin 4 and the following commands were used to create the table schema. Four tables were created, and the column names and value types were specified. The 'state' column was set as the primary key in each table because it is present in all tables.

```
 CREATE TABLE wage_2009(
         year INT NOT NULL,
         state VARCHAR PRIMARY KEY,
         high_wage FLOAT NOT NULL,
         low_wage FLOAT NOT NULL,
         high_2018 FLOAT NOT NULL,
         low_2018 FLOAT NOT NULL
);

 CREATE TABLE wage_2009(
         year INT NOT NULL,
         state VARCHAR PRIMARY KEY,
         high_wage FLOAT NOT NULL,
         low_wage FLOAT NOT NULL,
         high_2018 FLOAT NOT NULL,
         low_2018 FLOAT NOT NULL
);
```

```
 CREATE TABLE education(
     state VARCHAR PRIMARY KEY,
     education_percent FLOAT NOT NULL
);

CREATE TABLE peace(
     state VARCHAR PRIMARY KEY,
     peace_percent FLOAT NOT NULL
);

SELECT * FROM wage_2009;
SELECT * FROM wage_2010;
SELECT * FROM education;
SELECT * FROM peace;
```

EXTRACTION

- Our first dataset was retrieved from Kaggle.com. It displays the minimum wage "maximum and minimum" for each state in the United States during the years 1968 and 2017.
- Our second dataset was retrieved from data.world.com. It displays 4 sociological metrics for all 50 states during different years.

- The following command provides the source path for each csv file and stores it in a variable. Then the variable is called into a pandas DataFrame, and the first five rows and all the columns of the dataset are displayed.

```
MinWage_file = "../Resources/MinimumWage_Data.csv"
MinWage_df = pd.read_csv(MinWage_file, encoding = 'latin1’)
MinWage_df.head()

Metrics_file = "../Resources/Metrics_Data.csv”
Metrics_df = pd.read_csv(Metrics_file)
Metrics_df.head()
```

TRANSFORMATION

- Several transformations were made to each DataFrame in preparation to export them to the tables made in pgAdmin 4.
- The columns in the minimum wage DataFrame (and dataset) include 1. state, 2. year, 3. high value, or the highest value of the several minimum wage values found for one state during a single year, 4. low value, or the lowest value of the several minimum wage values found for one state during a single year, 5. high 2018 represents the 2018 equivalent in dollars for the high value 6. low 2018 represents the 2018 equivalent in dollars for the low value. The dataset file contains other columns, but we are only interested in the previously mentioned columns. Therefore, we create a new variable that holds the names for the columns of interest. Then a new DataFrame is created, which contains only the columns stored in the previous variable. A copy of the original DataFrame is made to create the second DataFrame to keep the original intact. The columns of this new DataFrame are renamed to the column names given in the SQL tables.

```
MinWage_cols = ["Year", "State", "High.Value", "Low.Value", "High.2018", "Low.2018"]
MinWage_df_transformed = MinWage_df[MinWage_cols].copy()
```

- Rename the column headers

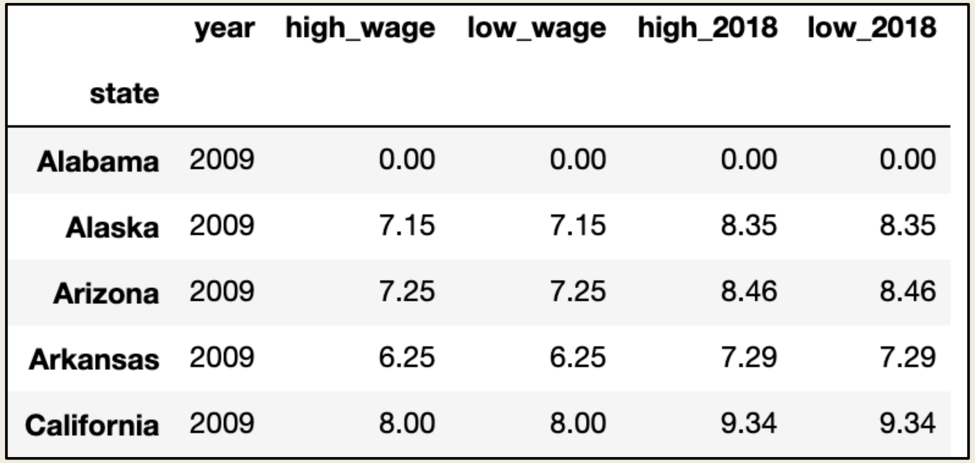
```
MinWage_df_transformed = MinWage_df_transformed.rename(columns={"Year":"year",
                                                                "State": "state",
                                                                "High.Value":"high_wage",
                                                                "Low.Value":"low_wage",
                                                                "High.2018":"high_2018",
                                                                "Low.2018":"low_2018”
                                                                 })
```

TRANSFORMATION

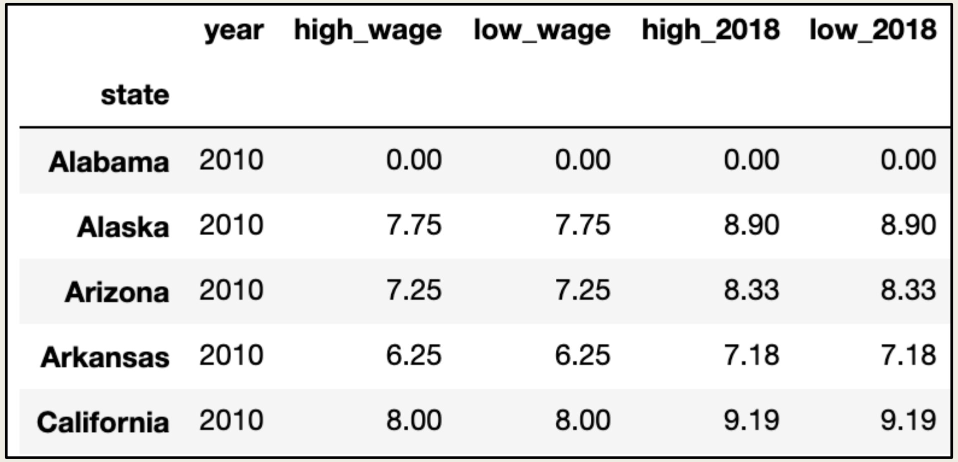
- From the DataFrame ‘MinWage_df’, two Dataframes were made by filtering by the years 2009 and 2010. We selected the rows where ‘Year’ = 2009 because the percent of educational attainment in the ‘Metrics_df’ table was collected during this year. Similarly, we selected the rows where ‘Year’ = 2010 because the percent peace index for each state in the ‘Metrics_df’ table was collected during this year. Each DataFrame was then grouped by state.

```
MinWage_2009 = MinWage_df_transformed.loc[MinWage_df_transformed['year'] == 2009]
MinWage_2009 = MinWage_2009.groupby('state').mean()
MinWage_2009.head()

MinWage_2010 = MinWage_df_transformed.loc[MinWage_df_transformed['year'] == 2010]
MinWage_2010 = MinWage_2010.groupby('state').mean()
MinWage_2010.head()
```



	year	high_wage	low_wage	high_2018	low_2018
state					
Alabama	2009	0.00	0.00	0.00	0.00
Alaska	2009	7.15	7.15	8.35	8.35
Arizona	2009	7.25	7.25	8.46	8.46
Arkansas	2009	6.25	6.25	7.29	7.29
California	2009	8.00	8.00	9.34	9.34



	year	high_wage	low_wage	high_2018	low_2018
state					
Alabama	2010	0.00	0.00	0.00	0.00
Alaska	2010	7.75	7.75	8.90	8.90
Arizona	2010	7.25	7.25	8.33	8.33
Arkansas	2010	6.25	6.25	7.18	7.18
California	2010	8.00	8.00	9.19	9.19

TRANSFORMATION

- The metrics dataset contains the following columns 1. state, 2. percentage of educational attainment, which displays the number of of individuals that have earned a bachelor's degree or higher during 2009, 3. percent peace index, which is measured based on homicide, violent crime, policing, incarceration and availability of small arms rates; data was converted to percentages and the higher the % the "more peaceful" the state, 4. above poverty rate, or the number of households living above poverty level, converted into %, and 5. percent non-religious is the % of individuals that do not identify as "highly religious.”

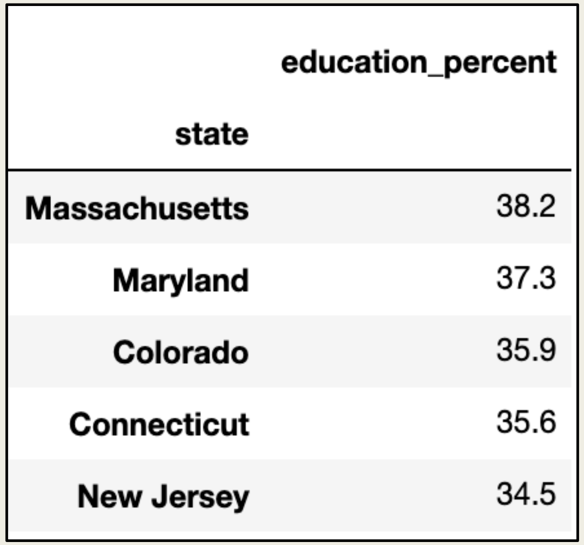
- For our first DataFrame, we extracted the percentage of educational attainment in 2009 and state columns. For our second DataFrame, we extracted the percent peace index in 2010 and state columns. The ‘sate’ column was set as the index, allowing all tables to be joined by this column.

TRANSFORMATION

```
MetricsEdu_cols = ["State", "Percent Educational Attainment"]
MetricsEdu_df = Metrics_df[MetricsEdu_cols].copy()

MetricsEdu_df = MetricsEdu_df.rename(columns={"State":"state",
              "Percent Educational Attainment": "education_percent"
                                                    })

MetricsEdu_df = MetricsEdu_df.set_index('state')
MetricsEdu_df.head()
```

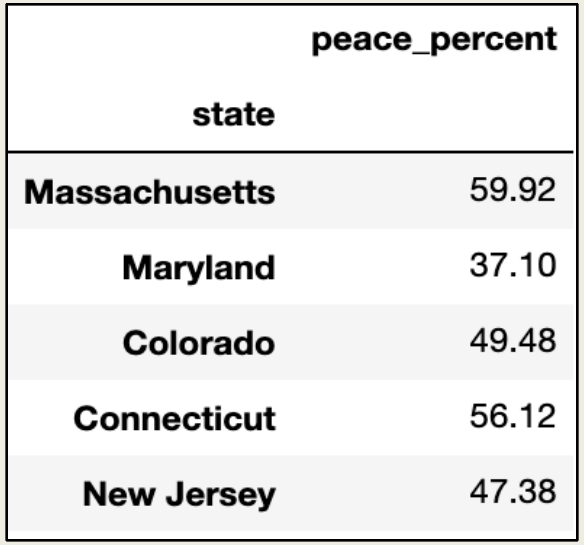


	education_percent
state	
Massachusetts	38.2
Maryland	37.3
Colorado	35.9
Connecticut	35.6
New Jersey	34.5

```
MetricsPeace_cols = ["State", "Percent Peace Index"]
MetricsPeace_df = Metrics_df[MetricsPeace_cols].copy()

MetricsPeace_df = MetricsPeace_df.rename(columns={"State": "state",
                                  "Percent Peace Index": "peace_percent"
                                                    })

MetricsPeace_df = MetricsPeace_df.set_index('state’)
MetricsPeace_df.head()
```



	peace_percent
state	
Massachusetts	59.92
Maryland	37.10
Colorado	49.48
Connecticut	56.12
New Jersey	47.38

LOAD

- A connection called ‘engine’ was created to connect to PostgresSQL pgAdmin 4, and the table column names was confirmed.

```
engine = create_engine('postgresql://postgres:postgres@localhost:5433/ETL_db’)
engine.table_names()
```

- The following commands load the DataFrames into the corresponding tables in pgAdmin 4, using the connection ‘engine’. The command if_exists='append’ adds the columns to an existing table or creates a new table if it does not already exist. The command ‘index=True’ means that we are supplying the index values to be populated in the primary key column.

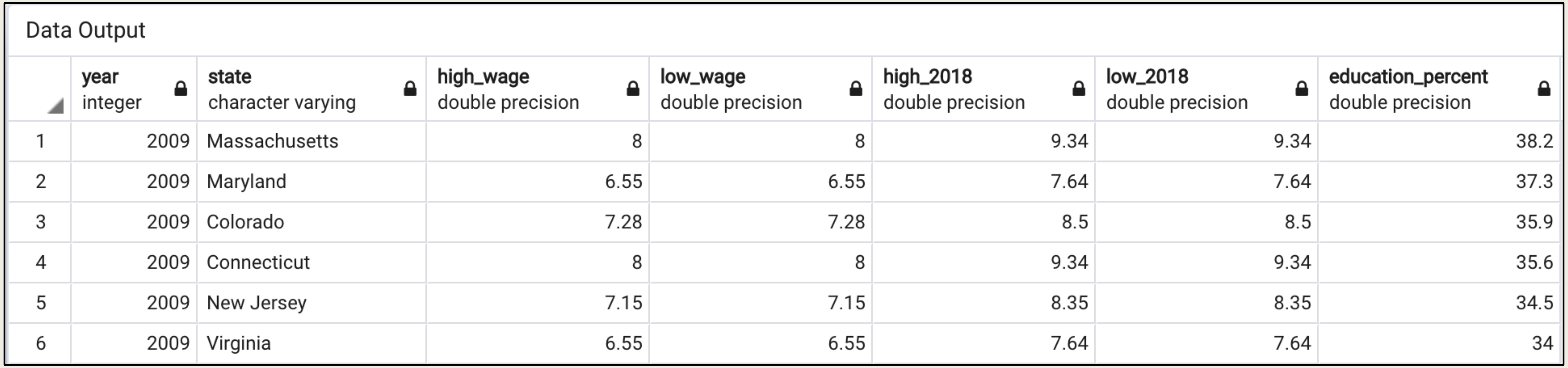
```
MinWage_2009.to_sql(name='wage_2009', con=engine, if_exists='append', index=True)
MinWage_2010.to_sql(name='wage_2010', con=engine, if_exists='append', index=True)
MetricsEdu_df.to_sql(name='education', con=engine, if_exists='append', index=True)
MetricsPeace_df.to_sql(name='peace', con=engine, if_exists='append', index=True)
```

SQL QUERY

- After the DataFrames were uploaded into the pgAdmin 4 tables, joined tables were made.

- The following query joins tables ‘wage_2009’ and ‘education’ at the primary key for each table, which is ‘state.’
- All the columns for table wage_2009 are shown, as well as column ‘education_percent’ from table education.

```
SELECT wage_2009.year,
wage_2009.state, wage_2009.high_wage,
wage_2009.low_wage,
wage_2009.high_2018,
wage_2009.low_2018,
education.education_percent
FROM wage_2009
JOIN education
ON wage_2009.state = education.state;
```

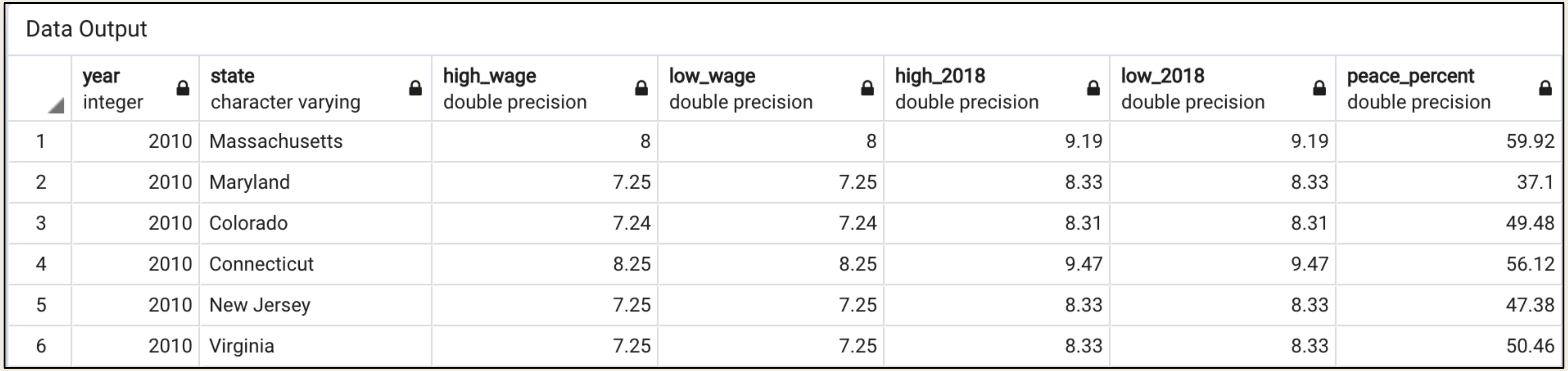


Data Output

	year integer	state character varying	high_wage double precision	low_wage double precision	high_2018 double precision	low_2018 double precision	education_percent double precision
1	2009	Massachusetts	8	8	9.34	9.34	38.2
2	2009	Maryland	6.55	6.55	7.64	7.64	37.3
3	2009	Colorado	7.28	7.28	8.5	8.5	35.9
4	2009	Connecticut	8	8	9.34	9.34	35.6
5	2009	New Jersey	7.15	7.15	8.35	8.35	34.5
6	2009	Virginia	6.55	6.55	7.64	7.64	34

SQL QUERY

```
SELECT wage_2010.year, wage_2010.state,
wage_2010.high_wage, wage_2010.low_wage,
wage_2010.high_2018, wage_2010.low_2018,
peace.peace_percent
FROM wage_2010
JOIN peace
ON wage_2010.state = peace.state;
```



Data Output

	year	state	high_wage	low_wage	high_2018	low_2018	peace_percent
	integer	character varying	double precision	double precision	double precision	double precision	double precision
1	2010	Massachusetts	8	8	9.19	9.19	59.92
2	2010	Maryland	7.25	7.25	8.33	8.33	37.1
3	2010	Colorado	7.24	7.24	8.31	8.31	49.48
4	2010	Connecticut	8.25	8.25	9.47	9.47	56.12
5	2010	New Jersey	7.25	7.25	8.33	8.33	47.38
6	2010	Virginia	7.25	7.25	8.33	8.33	50.46

HYPOTHESES

Hypothesis 1:

Minimum wage and educational attainment are directly proportional:

The higher the minimum wage, the higher the % of educational attainment.

The lower the minimum wage, the lower the % of educational attainment.

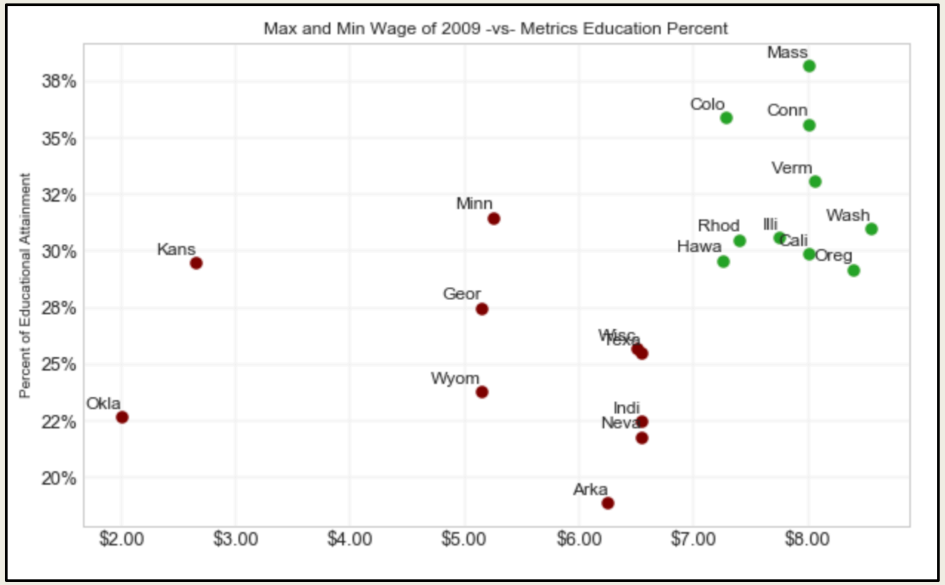
Hypothesis 2:

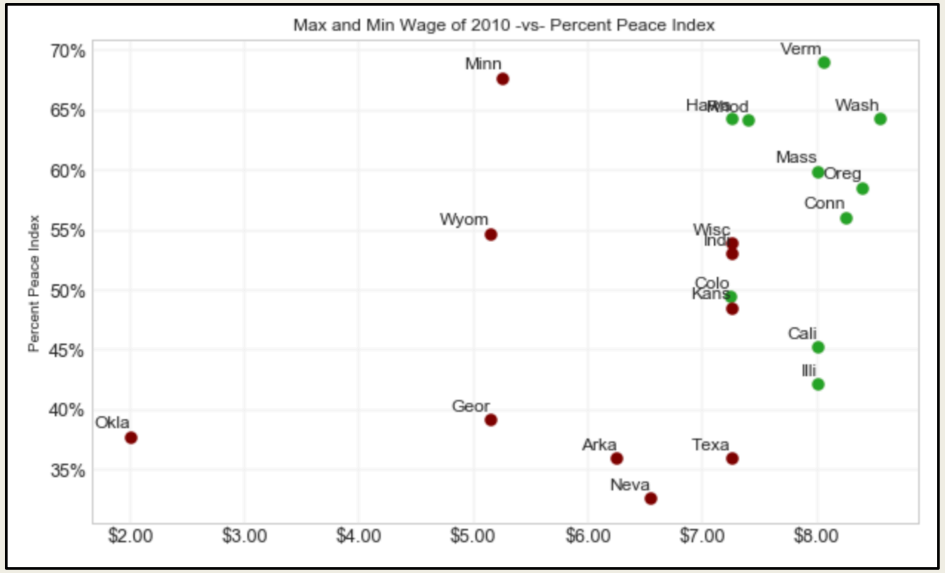
Minimum wage and % peace are directly proportional:

The higher the minimum wage, the higher the % peace.

The lower the minimum wage, the lower the % peace.

RESULTS





- Educational attainment and minimum wage were unrelated during the year 2009. The state with the highest minimum wage was Washington (8.55/hour, equivalent to 9.98/hour in 2018). However, this state was not the state with the highest % of education attainment during that same year. Moreover, the state with the lowest minimum wage value was Oklahoma (2.00/hour, equivalent to 2.33/hours in 2018). Similarly, this is not the state with the lowest % of educational attainment.

- Peace percent and minimum wage were unrelated during the year 2010. The state with the highest minimum wage was Washington (8.55/hour, equivalent to 9.82/hour in 2018). However, this state was not the state with the highest peace % during that same year. Moreover, the state with the lowest minimum wage value was Oklahoma (2.00/hour, equivalent to 2.30/hours in 2018). Similarly, this is not the state with the lowest peace %

REFERENCES

- Lislejoem. US Minimum Wage by State from 1968 to 2017 and 2018 Equivalent Dollars. *Kaggle.com*. Retrieved from https://www.kaggle.com/lislejoem/us-minimum-wage-by-state-from-1968-to-2017
- Nayar, K. US-States-Sociological-Metrics. Data.world. Retrieved from https://data.world/kevinnayar/us-states-sociological-metrics