Wage Analysis Across States and Demographics

Spring 2025

Nicholas Nemkov

Figure 1: Source: Unsplash

Introduction

How much money do you make?

A person's wage is influenced by various personal characteristics and broader state-level economic factors. State-level wage statistics provide insight into regional economic trends and job market conditions. I plan to examine and analyze wage statistics across different demographic groups, states, and professions.

The Data

```
import pandas as pd
import numpy as np
import matplotlib as plt
import plotly.express as px
import warnings
warnings.filterwarnings("ignore")

# load all datasets
adult_df = pd.read_csv('C:/Users/nickn/OneDrive/Desktop/Capstone Project/capstone/data/adult

# deal with NA values for adult.data
adult_df.dropna(inplace=True)

adult_df.columns = ["age", "workclass", "fnlwgt", "education", "education-num",
                    "marital-status", "occupation", "relationship", "race", "sex",
                    "capital-gain", "capital-loss", "hours-per-week", "native-country", "sa

state_wage_stats = pd.read_excel('C:/Users/nickn/OneDrive/Desktop/Capstone Project/capstone/d

population = pd.read_csv('C:/Users/nickn/OneDrive/Desktop/Capstone Project/capstone/data/NST-
```

I will be using three specific datasets:

1. adult.data

```
adult_df.head()
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	rel
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	No
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Hu
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	No
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Hu
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	W

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	rel

This dataset, extracted from the 1994 Census database, contains a 15 columns and 32561 observations. Each row represents a person, with 14 of their personal characteristics (age, sex, marital status, etc.) and a categorical target column stating whether their salary is less than or equal to $50k or greater than $50k. The majority of the employees are born in the United States, with others being from all around the world.

2. state_M2023_dl.xlsx

```
state_wage_stats.head()
```

	AREA	AREA_TITLE	AREA_TYPE	PRIM_STATE	NAICS	NAICS_TITLE	I_GROUP
0	1	Alabama	2	AL	0	Cross-industry	cross-industry
1	1	Alabama	2	AL	0	Cross-industry	cross-industry
2	1	Alabama	2	AL	0	Cross-industry	cross-industry
3	1	Alabama	2	AL	0	Cross-industry	cross-industry
4	1	Alabama	2	AL	0	Cross-industry	cross-industry

The *state_M2023_dl.xlsx* dataset is sourced from the US Bureau of Labor Statistics contains the 2023 occupational employment and wage estimates, calculated with data collected from employers in all industry sectors in metropolitan and nonmetropolitan areas in the US (including US territories). Each row represents an occupation title, its location in the US, and various wage statistic values (total employment, mean wage estimates, etc.).

3. NST-EST2023-ALLDATA.csv

```
# I will focus on just the state populations in 2023
population.head()
```

	SUMLEV	REGION	DIVISION	STATE	NAME	ESTIMATESBASE2020	POPESTIMA
0	10	0	0	0	United States	331464948	331526933
1	20	1	0	0	Northeast Region	57614141	57430477
2	30	1	1	0	New England	15119994	15057898
3	30	1	2	0	Middle Atlantic	42494147	42372579
4	20	2	0	0	Midwest Region	68987296	68969794

This U.S. Census Bureau dataset collects the estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States,

States, District of Columbia, and Puerto Rico from 2020 to 2023. Each row represents a US state and its yearly populations along with births and deaths.

> **!** Important
>
> I will be using the 2023 state populations only. This is later used in Figure 9 to calaculate the employment rate.

Research and Results

Now that all datasets are loaded, let's look at the wage statistics through a number of research questions:

Education Level Analysis

- Are longer weekly hour demanded on a job based on a level of education and greater age?
- Do higher-income workers tend to work longer hours?

```
# organize lists for education categories
preschool = [1]
primary = [2, 3]
secondary = [4, 5, 6, 7, 8, 9]
higher = [10, 11, 12, 13, 14, 15, 16]

def get_categories(number):
    '''
    Check if a numerical category is part of the 4 education groups.
    '''
    if number in preschool:
        return "Preschool"
    elif number in primary:
        return "Primary"
    elif number in secondary:
        return "Secondary"
    else:
        return "Higher"

# use get_categories function to create a new column with education categories
adult_df['education_category'] = adult_df['education-num'].apply(get_categories)
```

```
# create a decade column for easier viewing of the age
adult_df["age_decade"] = ((adult_df["age"]) // 10) * 10
mean_hours = adult_df.groupby(["age_decade", "education_category"])["hours-per-week"].mean()

# drop Preschool category if needed
drop_preschool = mean_hours[mean_hours['education_category'] != 'Preschool']

# Create line plot using Plotly Express
fig = px.line(drop_preschool, x="age_decade", y="hours-per-week", color="education_category"
              title="<b>Mean Hours Worked Per Week Against Employee Age</b>",
              labels={"age_decade": "Age", "hours-per-week": "Mean Hours Per Week", "educati
              markers=True)

# make sure text color is black
fig.update_layout(
    font=dict(
        color='black'
    ))

fig.show()
```

```
Unable to display output for mime type(s): text/html
```

(a)

```
Unable to display output for mime type(s): text/html
```

(b)

Figure 2

```
# boxplot of weekly hours worked against salary category
fig = px.box(adult_df, x="salary", y="hours-per-week", color="education_category",
             labels={
                     "education_category": "Education Level",
                     "salary": "Salary Category",
                     "hours-per-week": "Hours Worked per Week"
                 },
             title="Weekly Hours Against Salary Category")
fig.show()
```

Looking at Figure 2, we can observe that workers with Higher or Secondary educations tend

```
Unable to display output for mime type(s): text/html
```

Figure 3

to work longer weekly hours than those with only a Primary education. The interesting case with Primary education employees is that they work long weekly hours at age 10, and this decreases as they age increases.

Figure 3 shows us that employees work longer weeks for all education categories with a salary greater than $50k. This indicates that a higher salary job requires more time through the week.

Sex and Race Wage Trends

- Is there a noticable trend between a person's salary and their sex?
- Does a person's race influence his income?

```
# calculate total count of salary category by sex
counts = adult_df.groupby(['sex', 'salary']).size().reset_index(name='count')

# stacked bar chart
fig_2a = px.bar(counts, x='sex', y='count', color='salary',
                title="<b>Salary Levels Between Females and Males</b>")


# include black font for title
fig_2a.update_layout(
    font=dict(
        color='black'
    ))

fig_2a.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 4

Here in Figure 4 we can see for all men and women, significantly more men earn more than $50k as opposed to women (around 25% compared to 10%).

> **Note**
>
> This visual uses relatively old data (1994), it would be nice to compare a more modern visual to this one.

Now to see the spread by race:

```
# race occurences percentage
race_counts = adult_df.groupby(['race', 'salary']).size().reset_index(name='count')
race_counts['percentage'] = race_counts.groupby('race')['count'].transform(lambda x: (x / sum

fig_2d = px.bar(race_counts, x='race', y='percentage', color='salary',
                title="<b>Percentage of Salary Levels Between Races</b>")

fig_2d.update_layout(
    font=dict(
        color='black'
    ))

fig_2d.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 5

Figure 5 shows that the majority of the workers from each race category earn less than or equal to $50k. The White and Asian/Pacific Islander groups have the highest percentages of higher-salary (greater thn $50k) earners, over 25%. The other race groups have a lesser precentage of high-earners.

Salary Category Prediction

- What personal characteristics are predictive of a salary greater than $50,000?

The goal was to understand which personal characteristic affects a person's salary the most. I trained a logistic regression model to try to predict an employee's salary category (greater than $50,000 or less than or equal to $50,000) based on those characteristics (age, sex, marital status, education level, etc.).

Here is a full description of all features:

Feature	Description
age	Person's age.
workclass	Type of employment (private, self employed, etc.)
education	Education level
education-num	Education Numerical value.
marital-status	Marital status.
occupation	Type of job.
relationship	Familial status (Wife, Husband, Child, etc.)
race	Person's race.
sex	Person's sex.
capital-gain	Money gained from assets (investments, real-estate).
capital-loss	Money lost on your assets at the end of the year.
hours-per-week	Hours worked per week.
native-country	Country of birth.
fnlwgt	"Final Weight" for standardizing other characteristics.

Looking at the precision and recall from Figure 6:

Precision:

- When the model predicts a person to earn <=$50,000, it is correct 88% of the time.
- When predicting a salary greater than $50,000 The logistic regression model is 75% correct in all cases.

Recall:

- The model successfully identifies 93% of all actual <=$50,000 salary cases.
- Of all actual cases of greater than $50,000 salaries, 61% are correctly identified.

Let's also take a look at the absolute feature coefficients:

```
# Feature importance (absolute values)
feature_importance = np.abs(model.coef_[0])

importance_df = pd.DataFrame({"Feature": features.columns, "Importance": feature_importance})
importance_df = importance_df.sort_values(by="Importance", ascending=False)

# combine one-hot encoded features into the original feature
importance_df["Original Feature"] = importance_df["Feature"].str.split("_").str[0]
aggregated_importance = importance_df.groupby("Original Feature")["Importance"].sum()
aggregated_importance = aggregated_importance.sort_values(ascending=False)

# feature importance bar plot
```

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
# reference: https://realpython.com/logistic-regression-python/

# prepare adult_df data for regression

# split to features and target
features = adult_df.drop("salary", axis=1)
target = adult_df["salary"]

# one-hot encode categorical features
# reference: https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html
features = pd.get_dummies(features, drop_first=True)

# split data 70/30 train/test
features_train, features_test, target_train, target_test = train_test_split(features, target

# standardize numerical features
scaler = StandardScaler()
features_train = scaler.fit_transform(features_train)
features_test = scaler.transform(features_test)

# fit model
model = LogisticRegression()
model.fit(features_train, target_train)

predictions = model.predict(features_test)

print("Accuracy:\n", accuracy_score(predictions, target_test))
print("")
print("Classification Report:\n", classification_report(target_test, predictions))
```

```
Accuracy:
 0.850922753895458

Classification Report:
              precision    recall  f1-score   support

       <=50K       0.88      0.93      0.90      6767
        >50K       0.75      0.61      0.67      2282

    accuracy                           0.85      9049
   macro avg       0.81      0.77      0.79      9049
weighted avg       0.84      0.85      0.85      9049
```


Figure 6

```
features = aggregated_importance.sort_values(ascending=False).reset_index()
features.columns = ['Original Feature', 'Value']

# Create the bar plot
fig = px.bar(
    features,
    x='Value',
    y='Original Feature',
    orientation='h',
    title="<b>Absolute Feature Importances</b>",
    labels={
        "Original Feature": "Absolute Feature",
        "Value": "Coefficient Value"
    },
    category_orders={
        "Original Feature": features['Original Feature'].tolist()
    }
)

fig.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 7

As seen in Figure 7, the *capital-gain* feature greatly affects a person's salary. Capital gain is revenue that you recieve after selling a capital asset. This may include investments and real-estate (Investopedia Staff 2024). Other important features like *occupation* and *education* show us that a good education and important job title will also affect a person's salary. Features like *hour-per-week* and *capital-loss* have a much lesser effect on a person's salary.

State Average Salary Analysis

- Is there high wage variation among the US states in terms of mean hourly salary?
- Does geographic location influence income?

```
# filter state_wage_stats to only have rows with 'All Occupations' and contain US states
filt_state_df = state_wage_stats[state_wage_stats['OCC_TITLE'].str.contains('All Occupations
us_filt_state = filt_state_df[filt_state_df['AREA_TYPE'] == 2]
```

```
us_filt_state['A_MEAN'] = pd.to_numeric(us_filt_state['A_MEAN'])
us_filt_state['H_MEAN'] = pd.to_numeric(us_filt_state['H_MEAN'])
```

```
# for mean hourly wage
fig_4map_a = px.choropleth(
    us_filt_state,
    locations="PRIM_STATE",
    locationmode="USA-states",  # Use U.S. states
    color="H_MEAN",  # Value to color by
    color_continuous_scale="Reds",  # Color scale
    scope="usa",  # Focus on the USA
    title="Mean Hourly Salary By State",
    labels={"H_MEAN":"Hourly Wage",
            "PRIM_STATE":"State"}
)

fig_4map_a.update_layout(margin=dict(l=0, r=0, b=0, t=0),
                  width=800,
                  height=480)

fig_4map_a.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 8

The map (Figure 8) shows that states like New York and California have an average hourly pay of around $37. On the low end, the state of Mississippi has a mean hourly wage of only $22.87. The majority of the states have an hourly salary of slightly below $30. We can also see that the East Coast and West Coast have the higher salaries, probably due to those areas being high tech sectors.

Employment Rate Analysis

- Do high employment levels indicate high average salaries?

```
# prepare population dataset
state_population = population[population['SUMLEV'] == 40]
state_population = state_population[state_population['NAME'] != 'Puerto Rico']  # Drop all ar
state_population = state_population.rename(columns={'NAME': 'AREA_TITLE'})
```

```
pop_merge = pd.merge(us_filt_state, state_population, on='AREA_TITLE', how='left')

# create column with normalized employment rates
pop_merge['EMPLOYMENT_RATE'] = (pop_merge['TOT_EMP'] / pop_merge['POPESTIMATE2023'])
pop_merge['EMPLOYMENT_RATE'] = pd.to_numeric(pop_merge['EMPLOYMENT_RATE'])

pop_merge['EMPLOYMENT_RATE'] = pop_merge['EMPLOYMENT_RATE'].round(4)
```

First I want to visualize the employment rate by state:

```
fig_5a = px.choropleth(
    pop_merge,
    locations="PRIM_STATE",
    locationmode="USA-states",  # Use U.S. states
    color="EMPLOYMENT_RATE",
    color_continuous_scale="Viridis",
    scope="usa",
    title="Employment Rate By State",
    labels={"EMPLOYMENT_RATE":"Employment Rate",
            "PRIM_STATE":"State"}
)

fig_5a.update_layout(margin=dict(l=0, r=0, b=0, t=0),
                  width=800,
                  height=480)

fig_5a.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 9: Employment Rate by State

It can be noted in Figure 9 that for most states the employment rate is between 0.4 and 0.5. This indicates that of the total state population, 40% to 50% of people in that state are employed. This came as a surprise to me, as I expected a higher employment rate for the US.

```
# Annual Wages
A_filtered = pop_merge
A_filtered[['PRIM_STATE', 'TOT_EMP', 'A_MEAN', 'A_MEDIAN']].dropna()
```

```
# drop Washington DC, the outlier
A_filtered = A_filtered[A_filtered['AREA_TITLE'] != 'District of Columbia']

A_filtered['TOT_EMP'] = pd.to_numeric(A_filtered['TOT_EMP'], errors='coerce')
A_filtered['A_MEAN'] = pd.to_numeric(A_filtered['A_MEAN'], errors='coerce')
A_filtered['A_MEDIAN'] = pd.to_numeric(A_filtered['A_MEDIAN'], errors='coerce')

fig5_b = px.scatter(A_filtered,
                 x='EMPLOYMENT_RATE',
                 y='A_MEAN',
                 size='A_MEDIAN',  # Bubble size
                 hover_name='PRIM_STATE',
                 title="State Annual Wages vs. Employment",
                 labels={'TOT_EMP': 'Total Employment',
                         'A_MEAN': 'Annual Mean Wage',
                         'EMPLOYMENT_RATE': 'Employment Rate',
                         'A_MEDIAN': 'Annual Median Wage'},
                 size_max=20,  # Limits the max bubble size
                 color='A_MEDIAN'  # Color gradient based on wages
                )

fig5_b.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 10

The bubble chart in Figure 10 shows that as employment rate increases, so does the annual average salary. This may indicate that with a higher salary, an employee is more likely to stay in the position. The average median salary becomes greater as the employment rate increases.

This means that both high-earners and middle-level employees are benefitting in a state with a high employment rate.

IT Occupations Case Study

- In the case of IT professions, are there major salary differences between high-earners and middle-level employees?

```
# filter data to contain only IT-related professions
filt_state = state_wage_stats[state_wage_stats['AREA_TYPE'] == 2]

it_fields = ['computer', 'data', 'programmer', 'software', 'cyber', 'network']
pattern = '|'.join(it_fields)

it_job = filt_state[filt_state['OCC_TITLE'].str.contains(pattern, case=False, na=False)]

# list of non-IT jobs
drop_jobs = ['Computer and Mathematical Occupations',
             'Data Entry Keyers',
             'Computer Occupations, All Other',
             'Computer, Automated Teller, and Office Machine Repairers',
             'Electronics Engineers, Except Computer',
             'Office Machine Operators, Except Computer',
             'Computer Numerically Controlled Tool Operators',
             'Computer Hardware Engineers']

it_job = it_job[~it_job['OCC_TITLE'].isin(drop_jobs)]

it_job['A_PCT90'] = pd.to_numeric(it_job['A_PCT90'], errors='coerce')
it_job['A_MEDIAN'] = pd.to_numeric(it_job['A_MEDIAN'], errors='coerce')

# calculate difference between 90th percentile and median annual wage
it_job['difference_90_med'] = it_job['A_PCT90'] - it_job['A_MEDIAN']

# calculate the average difference for each state
state_avg_diff = it_job.groupby('PRIM_STATE')['difference_90_med'].mean().reset_index()
```

```
fig_6a = px.choropleth(
    state_avg_diff,
    locations= 'PRIM_STATE',
    locationmode="USA-states",
    color="difference_90_med",
    #color_continuous_scale="Viridis",
    scope="usa",
    title="Average Difference between 90th percentile and Median Annual Salaries for All IT J
    labels= {
        "difference_90_med": "Difference",
        "PRIM_STATE": "State"},
)
```

```
fig_6a.update_layout(margin=dict(l=0, r=0, b=0, t=0),
                  width=800,
                  height=480)

fig_6a.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 11: Average Difference between 90th percentile and Median Annual Salaries for IT Jobs

We can see from the above graph that New York and California IT professions have the highest divide in yearly salaries bewteen high and middle earners (well over $60,000 difference).

US State Profession Case Study

- Does a rural state like Nebraska share similar professions to urban states like New York and New Jersey?
- What is the most common occupation across the United States?

```
# filter data for only the three states
ny_nj_nb = ['NY', 'NJ', 'NE']
nynjnb_df = state_wage_stats[state_wage_stats['PRIM_STATE'].isin(ny_nj_nb)]
nynjnb_df = nynjnb_df[nynjnb_df['OCC_TITLE'] != 'All Occupations']
nynjnb_df = nynjnb_df[nynjnb_df['JOBS_1000'] != '**']
nynjnb_df['JOBS_1000'] = nynjnb_df['JOBS_1000'].astype(int)

# dataframe of top 5 jobs per state
top5_jobs_per_state = (
    nynjnb_df
    .sort_values(['PRIM_STATE', 'JOBS_1000'], ascending=[True, False])
    .groupby('PRIM_STATE')
    .head(5)
)
```

```
fig7_a = px.bar(
    top5_jobs_per_state,
    x='JOBS_1000',
    y='OCC_TITLE',
    color='PRIM_STATE',
    facet_col='PRIM_STATE',
```

```
    labels={'OCC_TITLE': 'Occupation', 'JOBS_1000': 'Jobs per 1000', 'PRIM_STATE': 'State'},
    height=450,
    orientation='h'
)

fig7_a.update_layout(showlegend=False)  # hide legend, not needed
fig7_a.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 12: Top 5 Occupations by Job per 1000 in NY, NJ, and NE

Referring to Figure 12, I can observe that Nebraska has the same 5 most common jobs as New Jersey. It came as a surprise that New York and New Jersey have different sets of common jobs. Given that the two states are so close, I though that their professions would be quite similar.

```
# filter data for most popular occupation per state
job_df = state_wage_stats[state_wage_stats['OCC_TITLE'] != 'All Occupations']
job_df = job_df[job_df['JOBS_1000'] != '**']
job_df['JOBS_1000'] = job_df['JOBS_1000'].astype(int)

top_job = (
    job_df
    .sort_values(['PRIM_STATE', 'JOBS_1000'], ascending=[True, False])
    .groupby('PRIM_STATE')
    .head(1)
)

fig7_b = px.choropleth(
    top_job,
    locations="PRIM_STATE",
    locationmode="USA-states",  # Use U.S. states
    color="OCC_TITLE",
    scope="usa",
    title="Most Popular Occupation Type By State",
    labels={"OCC_TITLE":"Occupation Type",
            "PRIM_STATE":"State"}
)

fig7_b.show()
```

```
Unable to display output for mime type(s): text/html
```

Figure 13

I also wanted to see if most US states share a common occupation type. After creating Figure 13, I can see that the majority of states have "Office and Administrative Support Occupations" as the most common type per 1000 jobs. There are four four that have different occupations, that being Nevada and Hawaii (Food Preparation and Serving Related Occupations), Indiana (Transportation and Material Moving Occupations), and Washington DC (Business and Financial Operations Occupations).

Conclusion

In summary, the project reveals several key takeaways:

- People with a higher education (completed highschool or recieved a college degree), tend to work longer hours than those with only a primary education.
- 25% of men have a salary greater than $50k, while around 10% of women have a salary greater than $50k.
- Capital gains will greatly increase a person's yearly revenue.
- New York IT jobs have the greatest difference between high-earner and middle-earner salaries.
- The most common jobs in the United States belong to the *Office and Administrative Support Occupations* category.

> **Note**
>
> Further information on this project can be found on my Github.

Becker, Barry, and Ronny Kohavi. 1996. "Adult." UCI Machine Learning Repository.

Investopedia Staff. 2024. "Capital Gain." https://www.investopedia.com/terms/c/capitalgain.asp.

U.S. Bureau of Labor Statistics. 2023. "May 2023 State Occupational Employment and Wage Estimates." https://www.bls.gov/oes/tables.htm.

U.S. Census Bureau, Population Division. 2023. "Annual Population Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, District of Columbia, and Puerto Rico: April 1, 2020 to July 1, 2023 (NST-EST2023-ALLDATA)." https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html#v2023.