Goals and a business objective:

- Create a marketing profile using US census bureau data, with a focus on the $50k salary benchmark.
- Evaluate how factors such as age, gender, education level, marital status, etc. can be used to determine if they earn above and below $50k.
- Boost enrollment for UVW college based on curated marketing profiles from the insights.

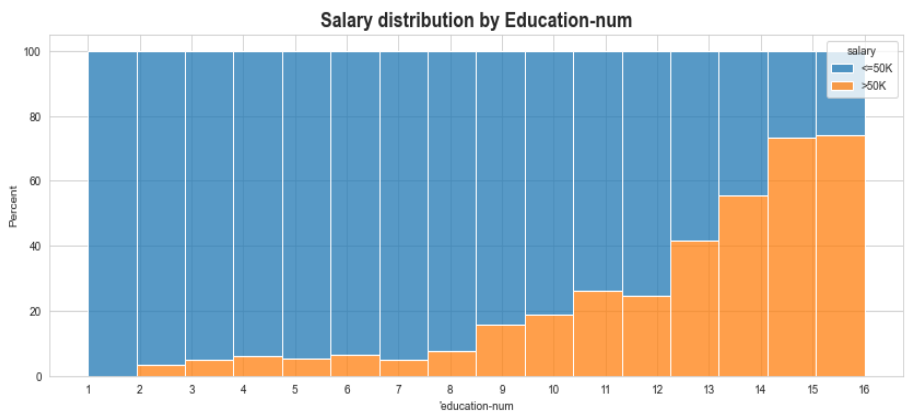
Assumptions:

- The audience can interpret a scatter plot, mosaic plot, parallel coordinate plot, etc.
- Data has been sampled appropriately and is accurate.
- Data has been grouped into its final state already and there is no need to categorize values within a feature (ex. dichotomize ‘work class’ feature into ‘public’ vs. ‘private’).

User Stories:

1. Marketing wants to understand how education-num influences income.
2. Marketing wants to understand how age influences income.
3. Marketing wants to understand how marital status influences income.
4. Marketing wants to understand how work class influences income.
5. Marketing wants to understand how race and sex influences income.
6. Marketing wants to understand how hours-per-week and native-country influence income.

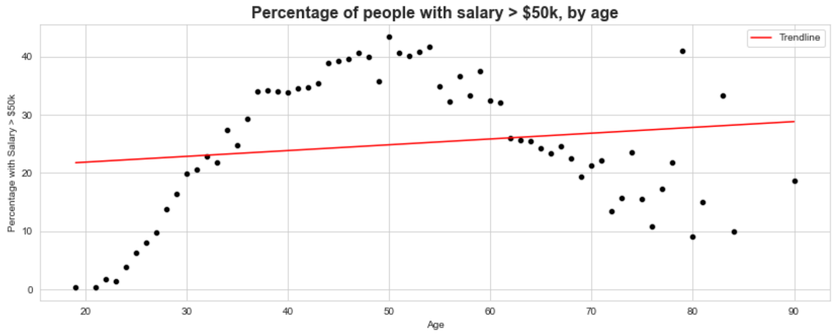
Visualizations:



User Story: Marketing wants to understand how education-num influences income.

Steps of the design process: Being a continuous numerical value, I chose a histplot to illustrate the distribution of values. From there I determined 16 bins would be most appropriate as that was the number of unique values in the field. I chose to show the values in percentage form, as the raw value counts skewed the data and made it hard to interpret the proportion of salary for each of the bins.

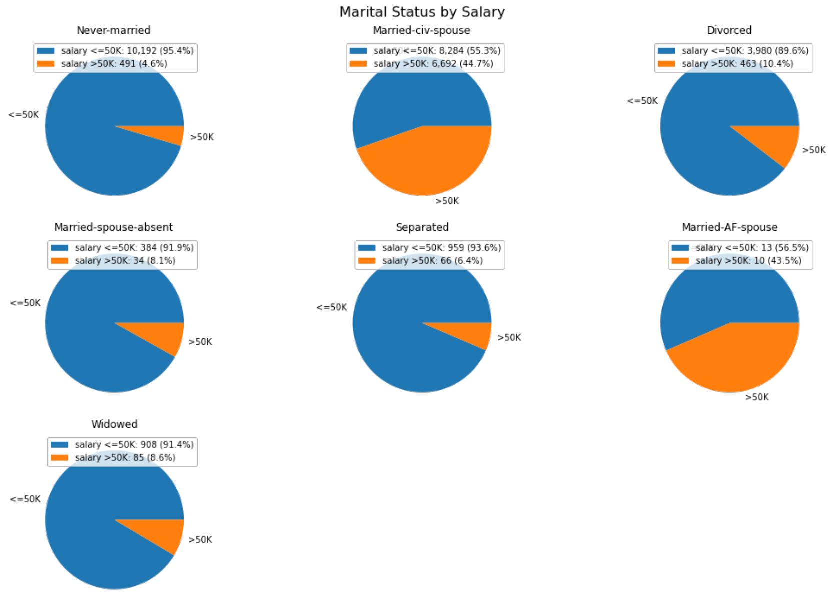
Conclusion: Education is positively correlated to income >$50k. The proportion of individuals with salary level >$50k is notably higher after obtaining level 13.



User Story: Marketing wants to understand how age influences income.

Steps of the design process: I wanted to show a scatter plot as I figured there might be some outliers with such a wide range of values. I chose to show as a percent so we can standardize the values and not have the visualization skewed by varying levels of observations. Showing as a percent also reduces clutter - there is only >$50k values plotted, but the <=$50k observations is also implied. To help illustrate the trend in the data I decided to plot a trendline that has been fit using linear regression.

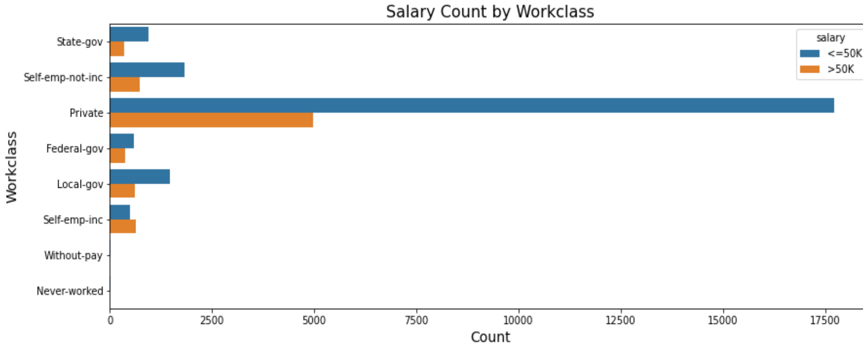
Conclusion: The highest proportion of earners >$50k is from ages 35-60, with a more prominent peak in the 45-55 age range. After age 70 the correlation between age and salary weakens. Age seems to be a strong indicator for salary.



User Story: Marketing wants to understand how marital status influences income.

Steps of the design process: This demonstrates the breakdown of salary for each marital status group. I initially decided to make 2 pie charts where 1 pie is <=$50k and the other is >$50k, with each pie divided by marital status. The problem with that was a category could have a drastically different proportion between pies but relatively even within itself. For example, married civ spouse was only 33% of the below pie but was 85% of the above. This initially led me to believe that this particular group was a good indicator of salary, but there are more observations of below than above. So I realized having just 2 pie charts was not going to work, and thus created a pie for each of the martial status groups, with the divide being salary.

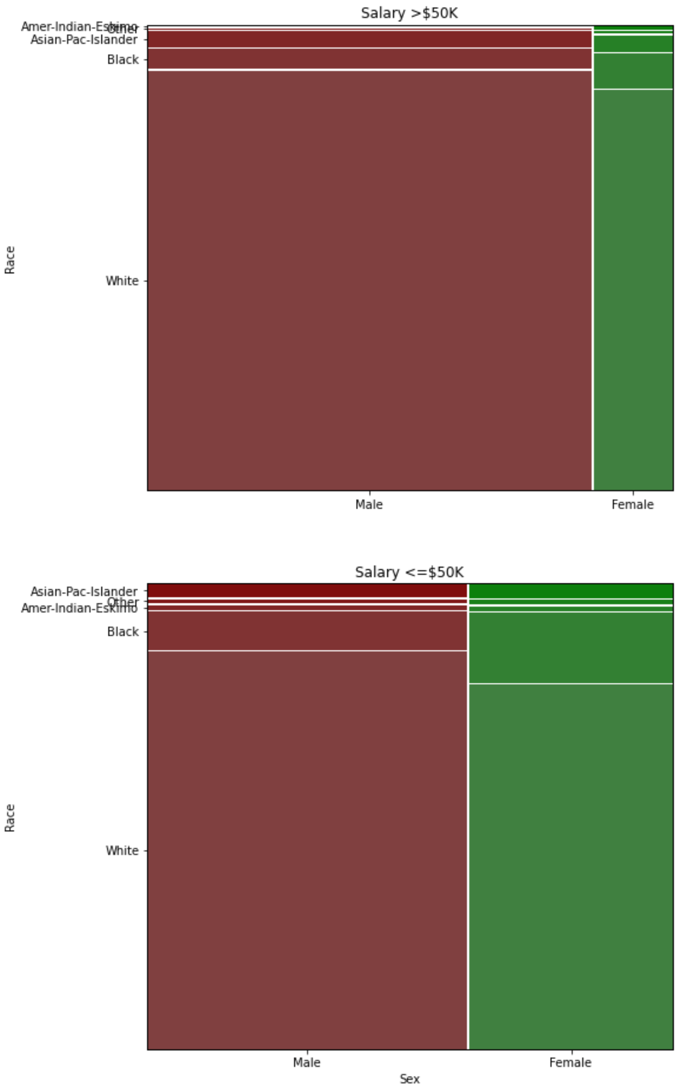
Conclusion: Many of the groups within this feature seem to be an indicator for income, so this feature in general looks to be a useful predictor.



User Story: Marketing wants to understand how work class influences income.

Steps of the design process: This demonstrates the occurrences of salaries below and above $50k for each of the groups in the feature ‘work class’. I chose this visualization because its useful to compare salary frequencies between each group in the dataset. The high observations of Private make it a good candidate for further examination. I chose to leave the two groups with 0 observations in the chart as it helps reaffirm the assumption that the data has integrity (it makes sense without pay and never worked have no salary observations).

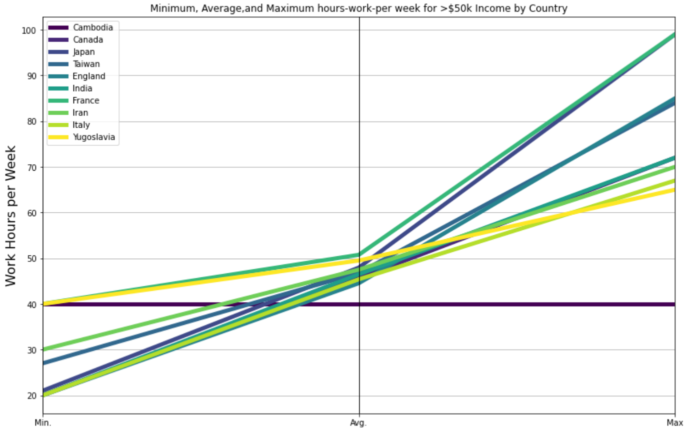
Conclusion: This feature appears to be a useful indicator for salary, in the sense that those self employed are more likely to earn $50k than earn less. Also, a state gov employee is less likely to earn >$50k than federal govt employee.

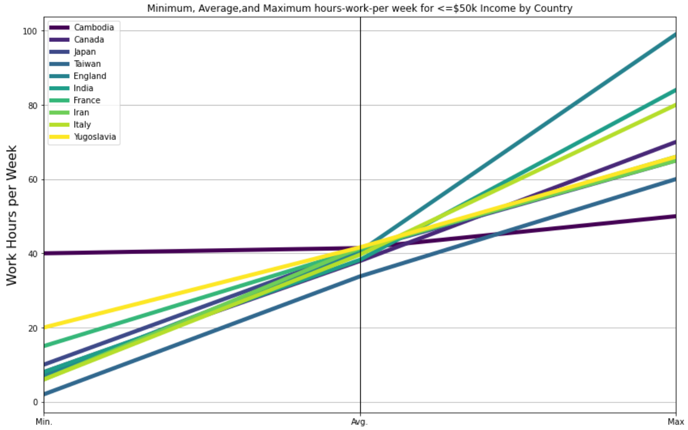


User Story: Marketing wants to understand how race and sex influences income.

Steps of the design process: There are just two values in the feature ‘sex’ which makes it a good candidate for display in the form of a mosaic plot. When combined with another categorical feature, race, we see some tiles with very little space which caused the labels to overlap initially. Therefore the design process was more about visual appeal rather than the story from the data as the story about the data was clear from the initial design. As we are trying to see how these features influence salary, there was need to make 2 mosaic plots to be able to compare to >$50k and <=$50k.

Conclusion: Sex seems to be a strong indicator, while race does as well. This is best evidenced by the white male tile increase from <=$50k to >$50k.





User Story: Marketing wants to understand how hours-per-week and native-country influence income.

Steps of the design process: What this demonstrates is the min, avg and max hours worked per week in countries where the proportion of those earning >$50k was in the 75th percentile. I limited the countries to make the visualization less crowded, and used this logic because this represents the country’s most capable of purchasing your product from an income perspective. There is a graph for >$50k and <=$50k so you can see the difference in a typical worker’s hours depending on the outcome and the country.

Conclusion: In countries within the 75th percentile, there are no >$50k earners working under 20hrs a week. On average within these countries >$50k earns work about 45-50hrs / week and about 38hrs for <=$50k.

Questions:

- I wasn’t sure if I should omit the values with small number of observations. For example, in the analysis of marital status as an indicator, there is a value of widowed that only has 12 occurrences in the set of 24k values. The low sample size may give way to skewed results and this needs to be included in the assumptions.

- Mosaic plot had a few values that had almost no representation on the visual. When the text labels were applied there was overlap. I initially increased the gap between tiles so they would not overlap, but then realized this is taking away from the presentation and the real story. In the end, labels on the axes were exclusively used.
- Is it better to aggregate groups where it seems appropriate? For example, in the work class feature, it might make sense to dichotomize the groups into ‘public’ and ‘private’, as ‘private’ exists already and has the majority of observations on its own. Ultimately, I thought this was outside of the scope of this exercise and so made just to make note in my assumptions above.

Not doing:

- I did not create a summary section aggregating my conclusions as this was not defined in the rubric. This however would have been helpful for the team creating the predictive model to be able to see the bottom-line analysis grouped together rather than spread throughout the report.
- I did not further investigate strong correlations between variables within certain groups. For example, marital status, work class, and race/sex were shown to be strong indicators of salary but only within certain groups. It would be interesting to see how to this analysis evolves with continued permutations of pairing with other features and slicing of groups.

Appendix:

- Please see the next page for an html rendering of the jupyter notebook containing the code behind these visualizations as well as more readable sized versions of these illustrations. It should be noted that the clarity and presentation of the code and reports is not as high as if viewing directly in jupyter notebooks, or an html file.

In [1]:

```
import pandas as pd
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.graphics.mosaicplot import mosaic
from scipy.stats import linregress
%matplotlib inline

df = pd.read_csv("adult.data.txt", header=None, sep=", ", engine='python')

df.columns = ["age", "workclass", "fnlwgt", "education", "education-num", "marital-
status", "occupation", "relationship", "race", "sex", "capital-gain", "capital-loss",
"hours-per-week", "native-country", "salary"]
for i in df.columns:
    df[i] = df[i].map(lambda x: None if str(x).strip() == '?' else x)


below = df[df["salary"] == "<=50K"]
above = df[df["salary"] == ">50K"]
```

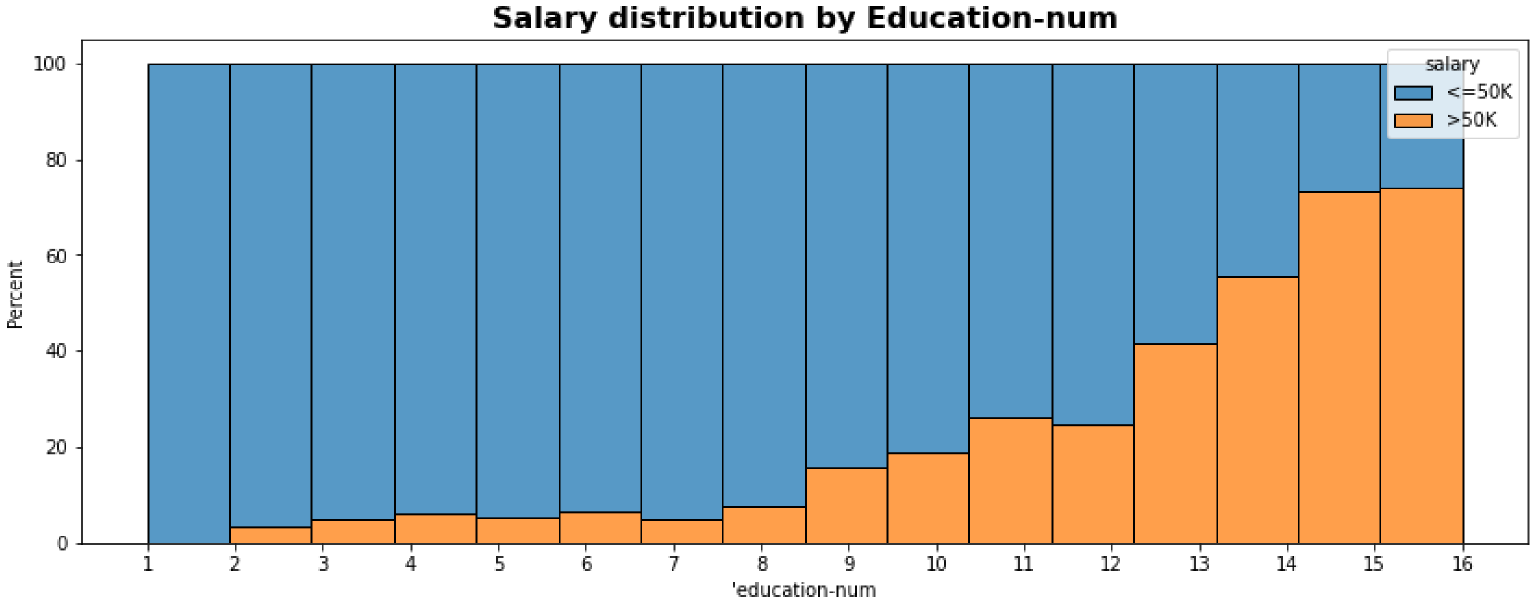
1. Marketing wants to understand how education-num influences income.

In [2]:

```
fig = plt.figure(figsize=(14,5))
plot = df.groupby(['education-num'])
['salary'].value_counts(normalize=True).mul(100).reset_index(name='percent')
g = sns.histplot(x='education-num', hue='salary', weights='percent', multiple='stack',
data=plot, bins=16)
g.set(ylabel="Percent", xlabel="'education-num")
g.set_xticks(np.arange(df['education-num'].min(), df['education-num'].max()+1, 1.0))
g.set_title("Salary distribution by Education-num", fontsize=16, fontweight='bold')
```

Out[2]:

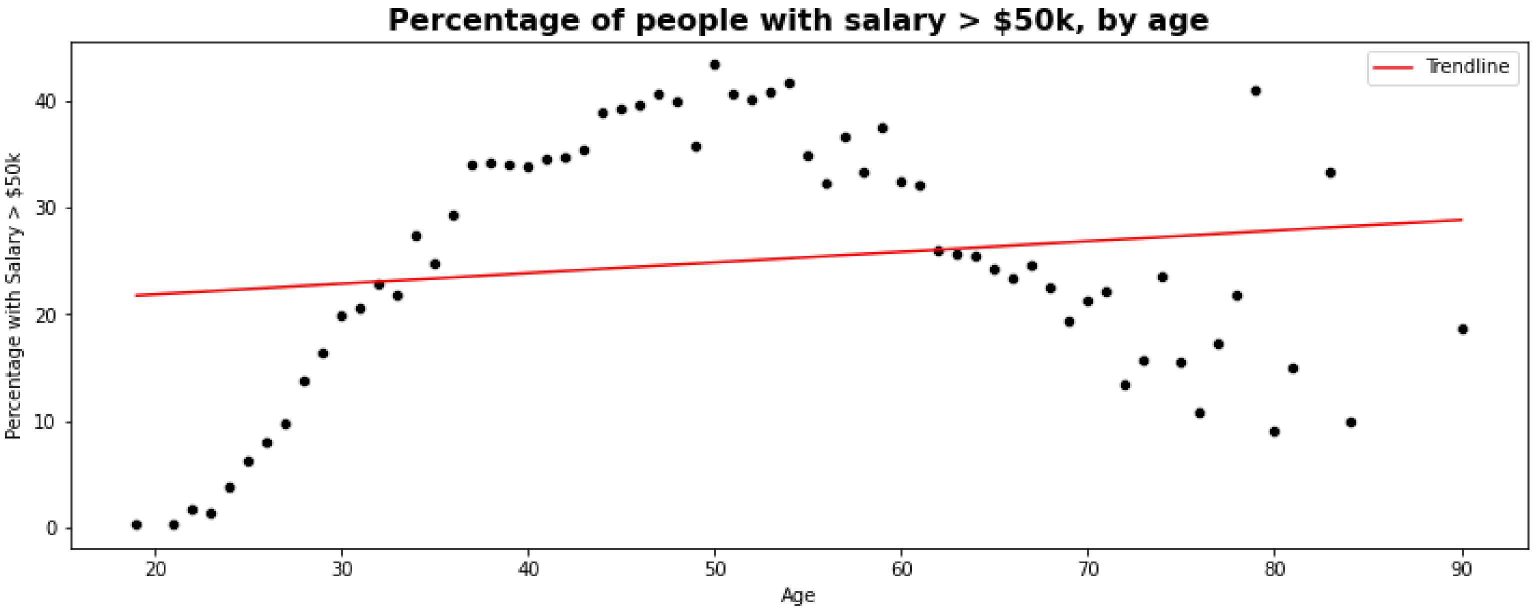
```
Text(0.5, 1.0, 'Salary distribution by Education-num')
```



1. Marketing wants to understand how age influences income.

In [3]:

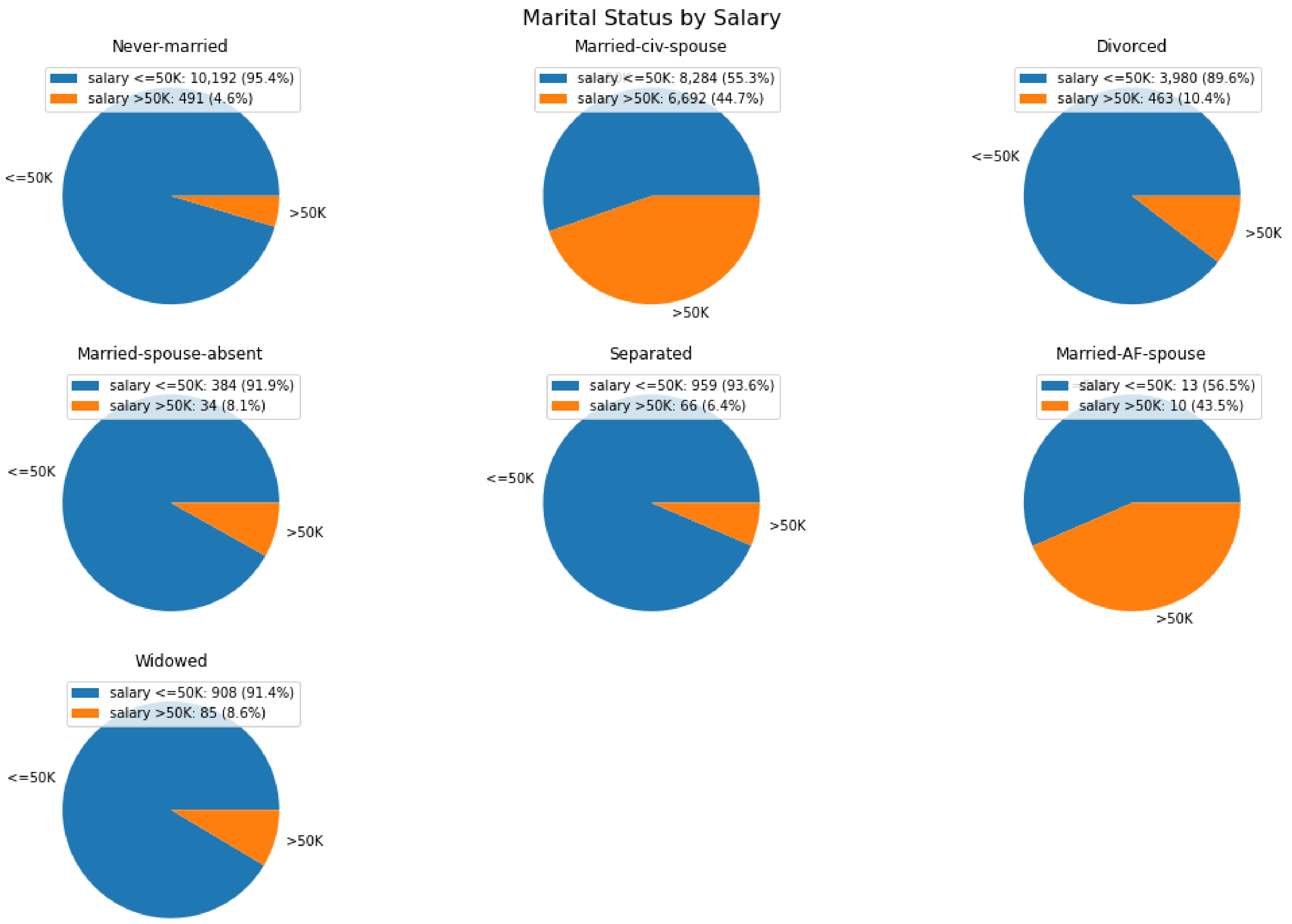
```
fig = plt.figure(figsize=(14,5))
plot = df.groupby(['age'])
['salary'].value_counts(normalize=True).mul(100).reset_index(name='percent')
df_above_50k = plot[plot['salary'] == '>50K']
axs = sns.scatterplot(x='age', y='percent', data=df_above_50k, color='black')
axs.set(xlabel='Age', ylabel='Percentage with Salary > $50k')
axs.set_title("Percentage of people with salary > $50k, by age", fontsize=16,
fontweight='bold')

slope, intercept, _, _, _ = linregress(df_above_50k['age'], df_above_50k['percent'])
trendline = slope * df_above_50k['age'] + intercept
plt.plot(df_above_50k['age'], trendline, color='red', label='Trendline')
plt.legend()
plt.show()
```



1. Marketing wants to understand how marital status influences income.

In [4]:

```
fig, axs = plt.subplots(nrows=3, ncols=3, figsize=(15, 10))

for i, metric in enumerate(df["marital-status"].unique()):
    ax = axs[i // 3, i % 3]
    raw = df[df["marital-status"] == metric]
    filtered = raw.groupby('salary').size().sort_values(ascending=False)
    filtered.plot(kind='pie', autopct='', ax=ax, title=f'{metric}')
    labels = [f'salary {i}: {format(v, ",.0f")} ({round((v/filtered.sum())*100,1)}%)'
for i,v in filtered.iteritems()]
    ax.legend(labels=labels, loc="best")
    ax.set_ylabel("")

plt.suptitle('Marital Status by Salary', fontsize=16)
axs[2, 2].set_visible(False)
axs[2, 1].set_visible(False)
plt.tight_layout()
plt.show()
```



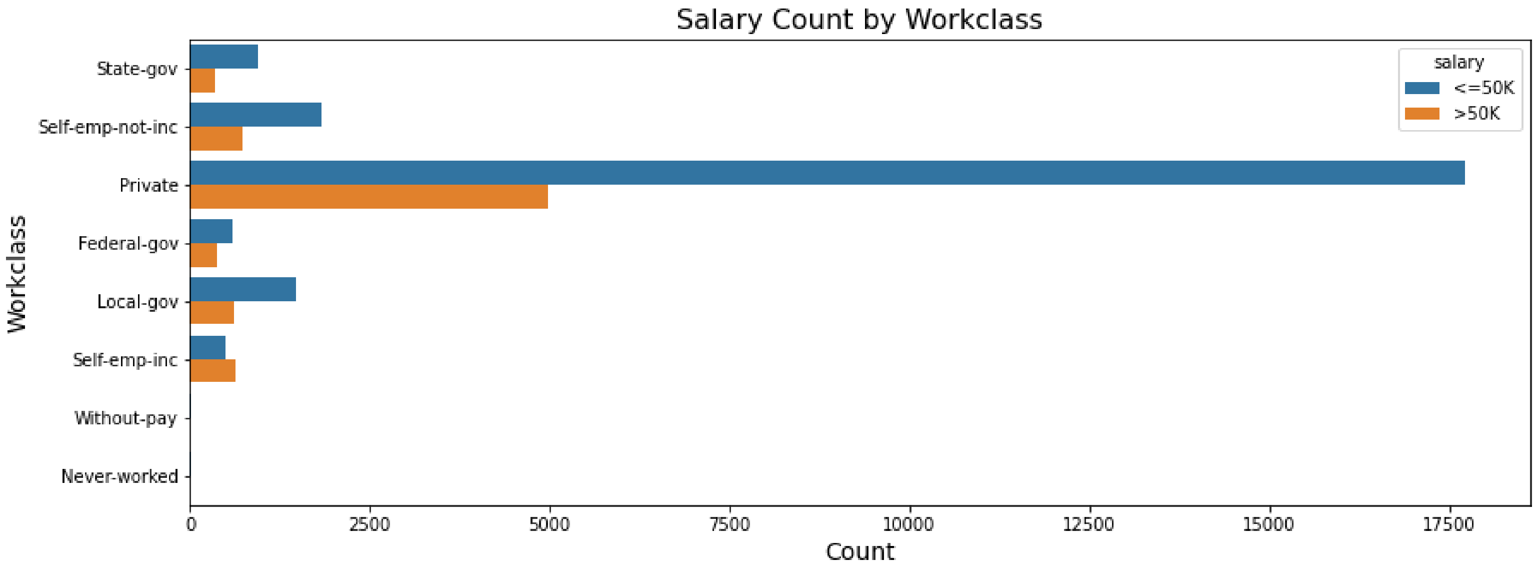
1. Marketing wants to understand how work class influences income.

In [5]:

```
fig = plt.figure(figsize=(14,5))
sns.countplot(data=df, y='workclass', hue='salary')
plt.title('Salary Count by Workclass',fontsize=16)
plt.xlabel('Count',fontsize=14)
plt.ylabel('Workclass',fontsize=14)
```

Out[5]:

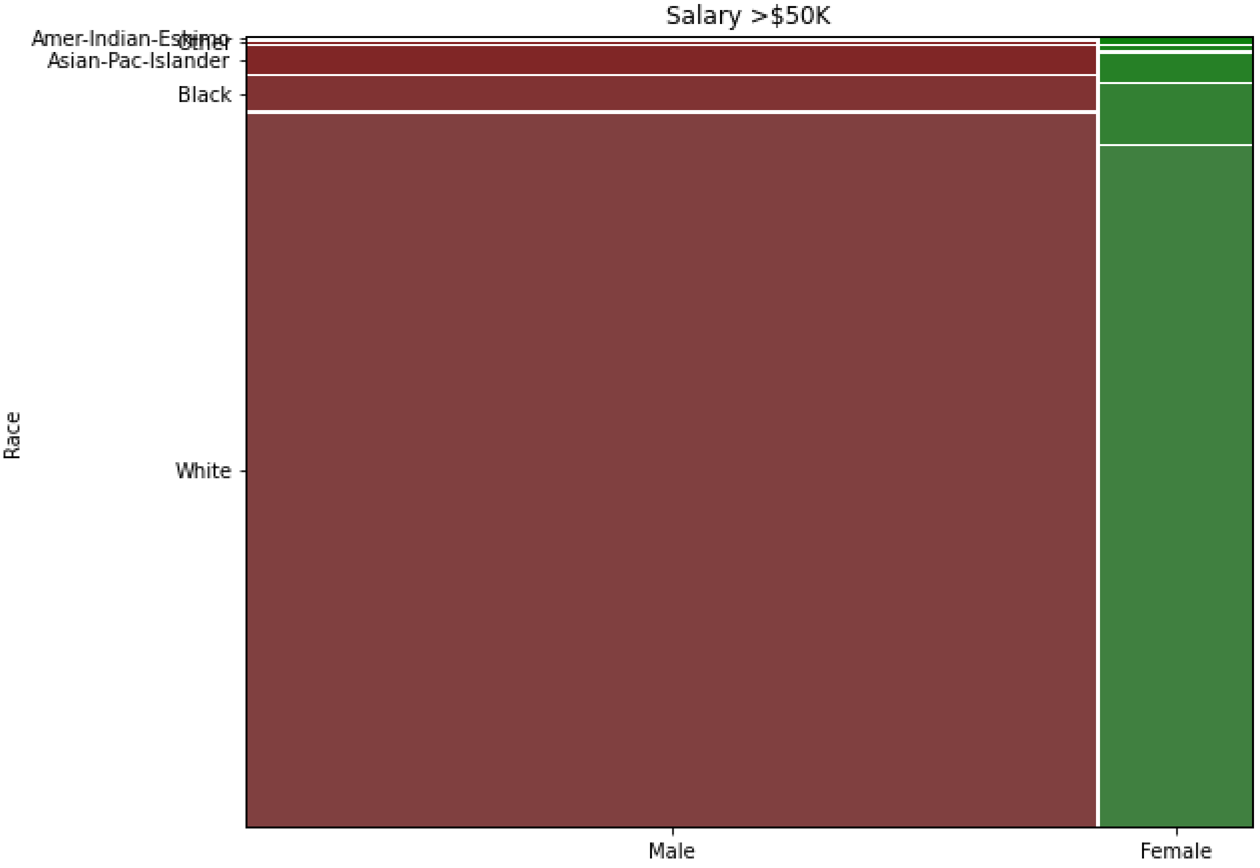
```
Text(0, 0.5, 'Workclass')
```

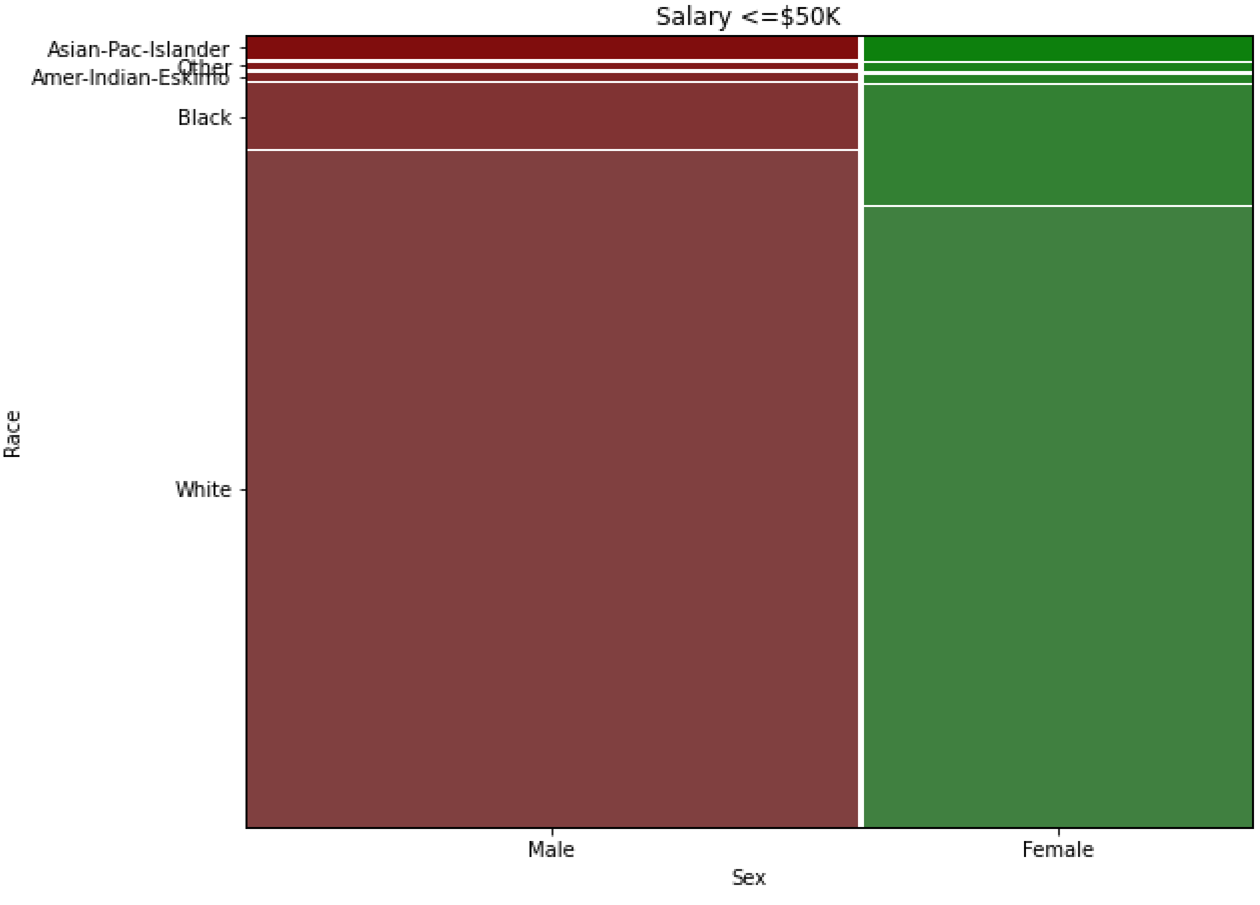


1. Marketing wants to understand how race and sex influences income.

In [9]:

```
plt.figure(figsize=(9,16))
ax1 = plt.subplot(211)
axes_dict = mosaic(above, ['sex', 'race'], ax=ax1, labelizer=lambda k: '')
ax1.set_title('Salary >$50K')

ax2 = plt.subplot(212)
mosaic(below, ['sex', 'race'], ax=ax2, labelizer=lambda k: '')
ax2.set_title('Salary <=$50K')
ax1.set_ylabel('Race')
ax2.set_ylabel('Race')
ax2.set_xlabel('Sex')
plt.show()
```

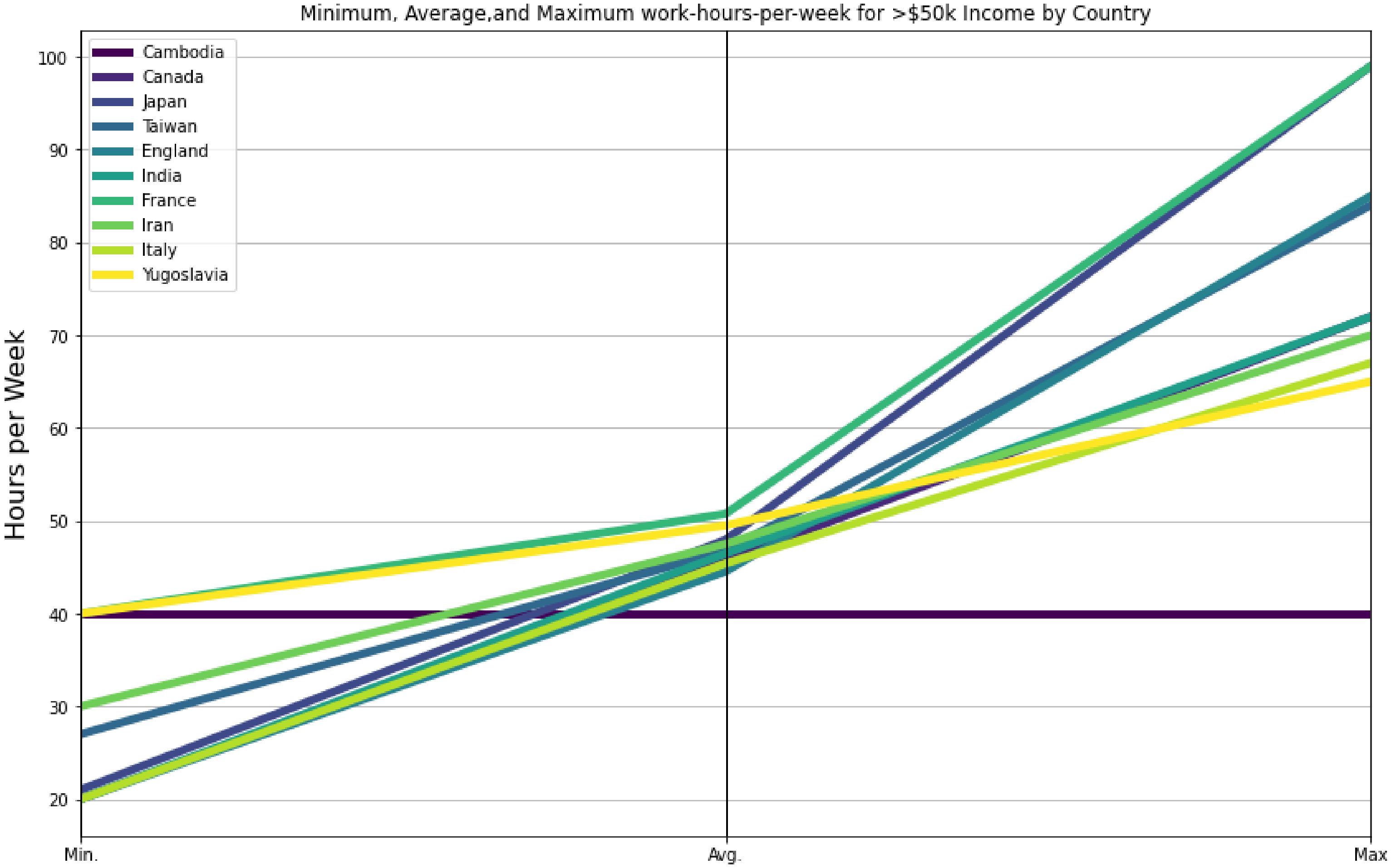


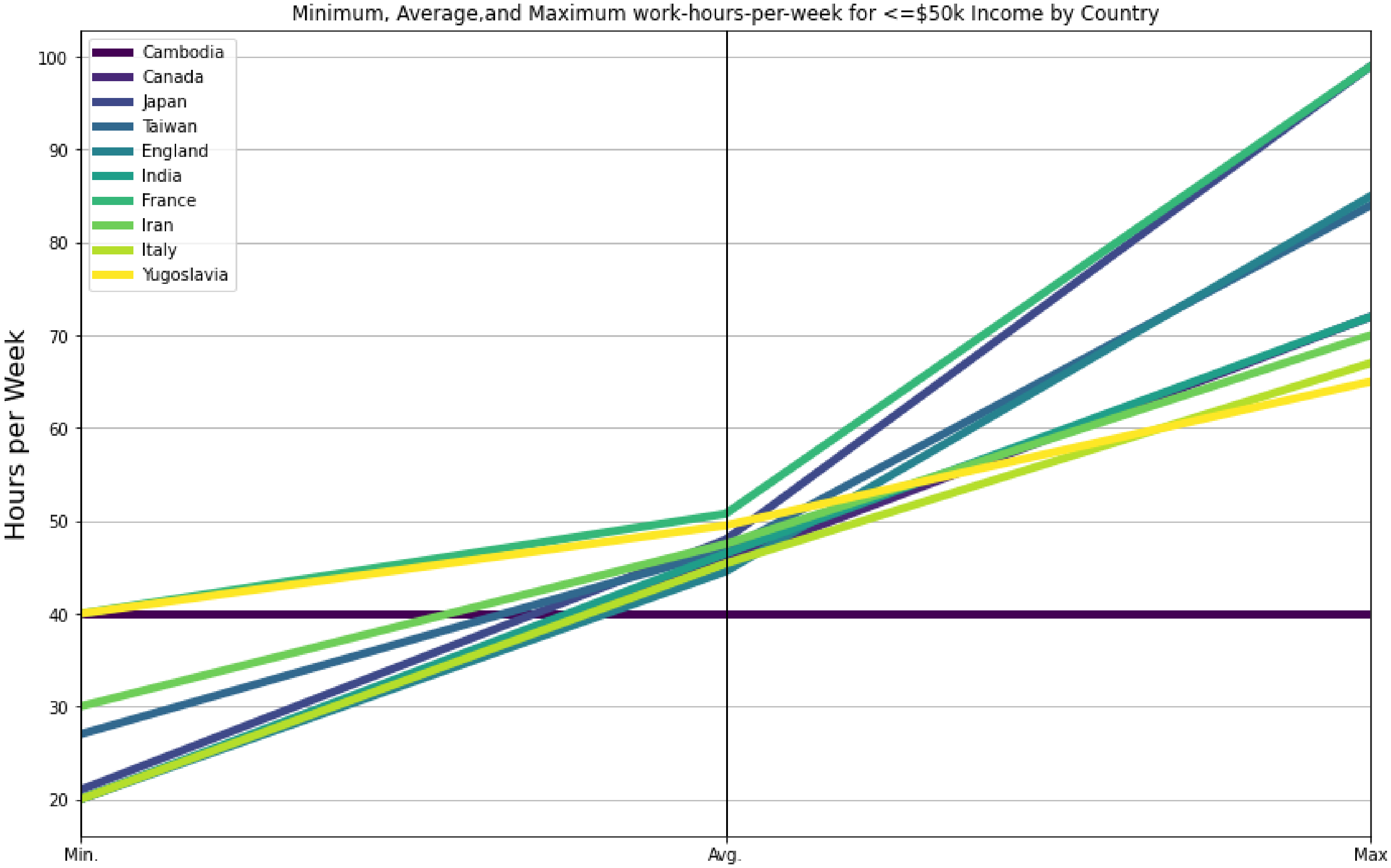


1. Marketing wants to understand how hours-per-week and native-country influence income.

In [8]:

```
dataabove = []
databelow = []

percents_above = dict(above.value_counts("native-country") /
(above.value_counts("native-country") + below.value_counts("native-country")))
mean = (above.value_counts("native-country") / (above.value_counts("native-country") +
below.value_counts("native-country"))).describe()["75%"]
countries_with_high_percent_of_above = sorted({k:i for k,i in percents_above.items()
if i > mean}, key=lambda x: x[1])

# countries = df_above_50k["native-country"].unique()
for country in countries_with_high_percent_of_above:

    min_attendance = min(above.loc[above['native-country'] == country]["hours-per-
week"].values)
    max_attendance = max(above.loc[above['native-country'] == country]["hours-per-
week"].values)
    avg_attendance = (sum(above.loc[above['native-country'] == country]["hours-per-
week"].values) /  len(above.loc[above['native-country'] == country]["hours-per-
week"].values))
    dataabove.append((country, min_attendance, avg_attendance, max_attendance))

    min_attendance = min(above.loc[above['native-country'] == country]["hours-per-
week"].values)
    max_attendance = max(above.loc[above['native-country'] == country]["hours-per-
week"].values)
    avg_attendance = (sum(above.loc[above['native-country'] == country]["hours-per-
week"].values) /  len(above.loc[above['native-country'] == country]["hours-per-
week"].values))
    databelow.append((country, min_attendance, avg_attendance, max_attendance))

plotdf = pd.DataFrame(dataabove, columns=["country", "Min.", "Avg.", "Max"])
fig, ax = plt.subplots(figsize=(14, 9))
pd.plotting.parallel_coordinates(plotdf, "country", linewidth=5, colormap="viridis")
plt.title('Minimum, Average,and Maximum work-hours-per-week for >$50k Income by
Country')
plt.ylabel("Hours per Week", fontsize=16)
plt.legend()
plt.show()

plotdf = pd.DataFrame(databelow, columns=["country", "Min.", "Avg.", "Max"])
fig, ax = plt.subplots(figsize=(14, 9))
pd.plotting.parallel_coordinates(plotdf, "country", linewidth=5, colormap="viridis")
plt.title('Minimum, Average,and Maximum work-hours-per-week for <=$50k Income by
Country')
plt.ylabel("Hours per Week", fontsize=16)
plt.legend()
plt.show()
```


Minimum, Average,and Maximum work-hours-per-week for >$50k Income by Country
Cambodia
Canada
Japan
Taiwan
England
India
France
Iran
Italy
Yugoslavia
Hours per Week
100
90
80
70
60
50
40
30
20
Min.
Avg.
Max



Minimum, Average,and Maximum work-hours-per-week for <=$50k Income by Country
Cambodia
Canada
Japan
Taiwan
England
India
France
Iran
Italy
Yugoslavia
Hours per Week
100
90
80
70
60
50
40
30
20
Min.
Avg.
Max