5.4.4 Create Box-and-Whisker Plots

You're feeling pretty good about your upcoming presentation. You reconvene with Sasha to thank her for her advice and see what final tips she might have. She thinks it would be a good idea to show V. Isualize if there are any outliers in the data. To do this, she suggests using box-and-whisker plots.

Up until now, we have gained experience creating a variety of charts where each chart can showcase the data in different ways, and we have performed summary statistics on the number of rides, the fare, and the number of drivers for each city type. Now, we are going to visualize the summary statistics and determine if there are any outliers by using box-and-whisker plots.

REWIND

Box-and-whisker plots are an effective way to show a lot of information about distribution in a small amount of space, especially outliers.

Box-and-Whisker Plots for Ride Count Data

Creating a box-and-whisker plot requires that we use the `ax.boxplot()` function, which takes an array inside the parentheses. We can also add a title and axes labels as we have done before.

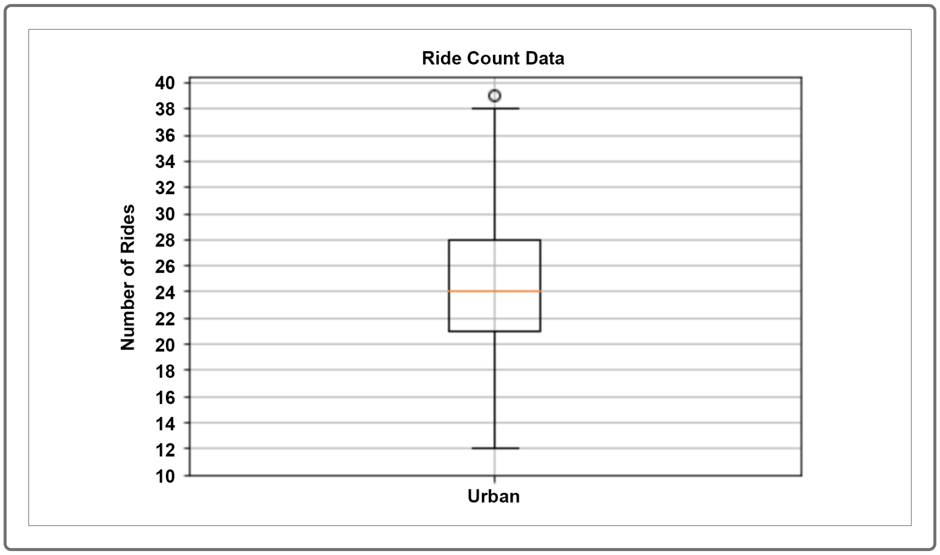
Let's create our `urban_ride_count` box-and-whisker plot. In a new cell, add the following code:

```
# Create a box-and-whisker plot for the urban cities ride count.
x_labels = ["Urban"]
fig, ax = plt.subplots()
ax.boxplot(urban_ride_count, labels=x_labels)
# Add the title, y-axis label and grid.
ax.set_title('Ride Count Data (2019)')
ax.set_ylabel('Number of Rides')
ax.set_yticks(np.arange(10, 41, step=2.0))
ax.grid()
plt.show()
```

Some of this code looks familiar, but a few lines are new. Let's break down what the code is doing.

- First, we create the x-axis labels with a list, `x_labels = ["Urban"]`.
- Next, the data and labels are passed in the `boxplot` function.
- Finally, we set the `y_ticks` with a range from 10 to 41 with ticks at an increment of 2. This will help determine where the minimum and maximum lie as well as any outliers.

When you run the cell, the urban ride count data box-and-whisker plot will look like this:

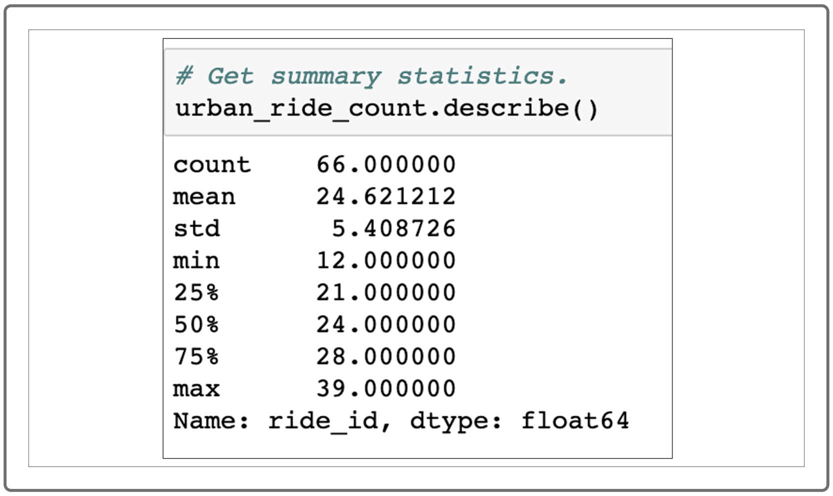


Looking at this box-and-whisker plot, we can see:

1. There is at least one outlier, which is close to 40. This our maximum data point, 39.
2. The minimum is 12.
3. The median is 24 or the 50th percentile.
4. The standard deviation is about 5 because the box upper and lower boundaries represent the upper and lower quartiles.

REWIND

We generated the same summary the box-and-whisker plot visual shows us by getting the high-level summary statistics using `urban_ride_count.describe()`.



```
# Get summary statistics.
urban_ride_count.describe()
```

```
count    66.000000
mean     24.621212
std       5.408726
min      12.000000
25%      21.000000
50%      24.000000
75%      28.000000
max      39.000000
Name: ride_id, dtype: float64
```

SKILL DRILL

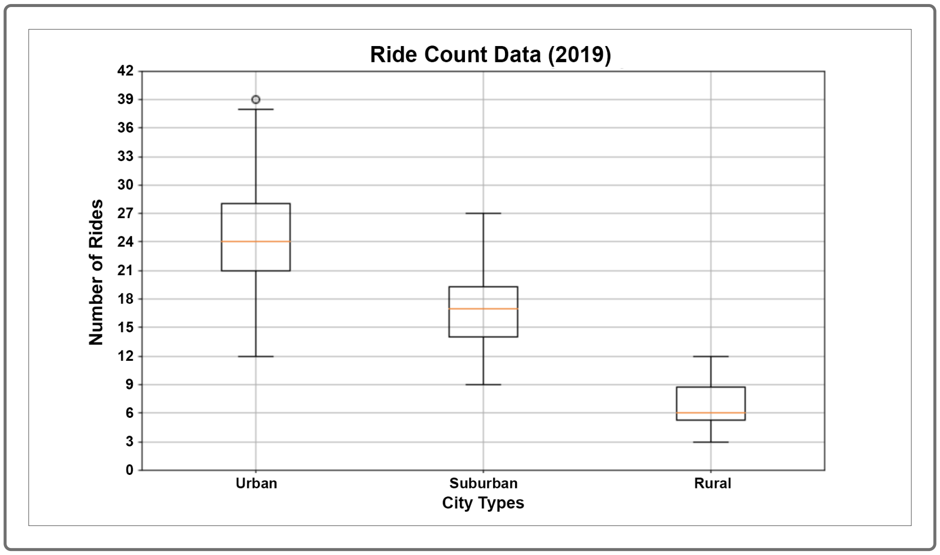
Calculate the summary statistics with box-and-whisker plots on the number of suburban and rural rides.

To show all the city type box-and-whisker plots on one chart we need to modify the `boxplot()` function and add other features. We will increase the size of the chart and the font of the title and axes labels.

Add the following code to the new cell and run the cell.

```
# Add all ride count box-and-whisker plots to the same graph.
x_labels = ["Urban", "Suburban","Rural"]
ride_count_data = [urban_ride_count, suburban_ride_count, rural_ride_count]
fig, ax = plt.subplots(figsize=(10, 6))
ax.set_title('Ride Count Data (2019)',fontsize=20)
ax.set_ylabel('Number of Rides',fontsize=14)
ax.set_xlabel("City Types",fontsize=14)
ax.boxplot(ride_count_data, labels=x_labels)
ax.set_yticks(np.arange(0, 45, step=3.0))
ax.grid()
# Save the figure.
plt.savefig("analysis/Fig2.png")
plt.show()
```

When we run this cell, the box-and-whisker plot below has all three individual box-and-whisker plots with the city type on the x-axis.



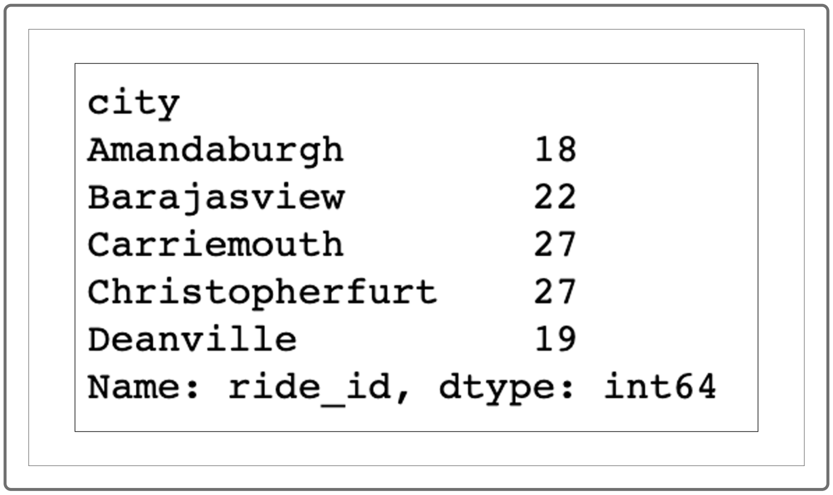
FINDING

There is one outlier in the urban ride count data. Also, the average number of rides in the rural cities is about 4- and 3.5-times lower per city than the urban and suburban cities, respectively.

One of our tasks was to find out if there were any outliers. We know that the outlier for the `urban_ride_count` is 39. From this information, we can find out which city has the highest rider count.

REWIND

Recall that the `urban_ride_count` is a Series with the index of the city and the data the number of rides for each city.



```
city
Amandaburgh          18
Barajasview          22
Carriemouth          27
Christopherfurt      27
Deanville            19
Name: ride_id, dtype: int64
```

We can get all the "True" values where the `urban_ride_count` equals 39. Then, we can filter the `urban_ride_count` Series for all the "True" values and get the city name from the index, like this:

```
# Get the city that matches 39.
urban_city_outlier = urban_ride_count[urban_ride_count==39].index[0]
print(f"{urban_city_outlier} has the highest rider count.")
```

The output from running this cell is `West Angela has the highest rider count.`

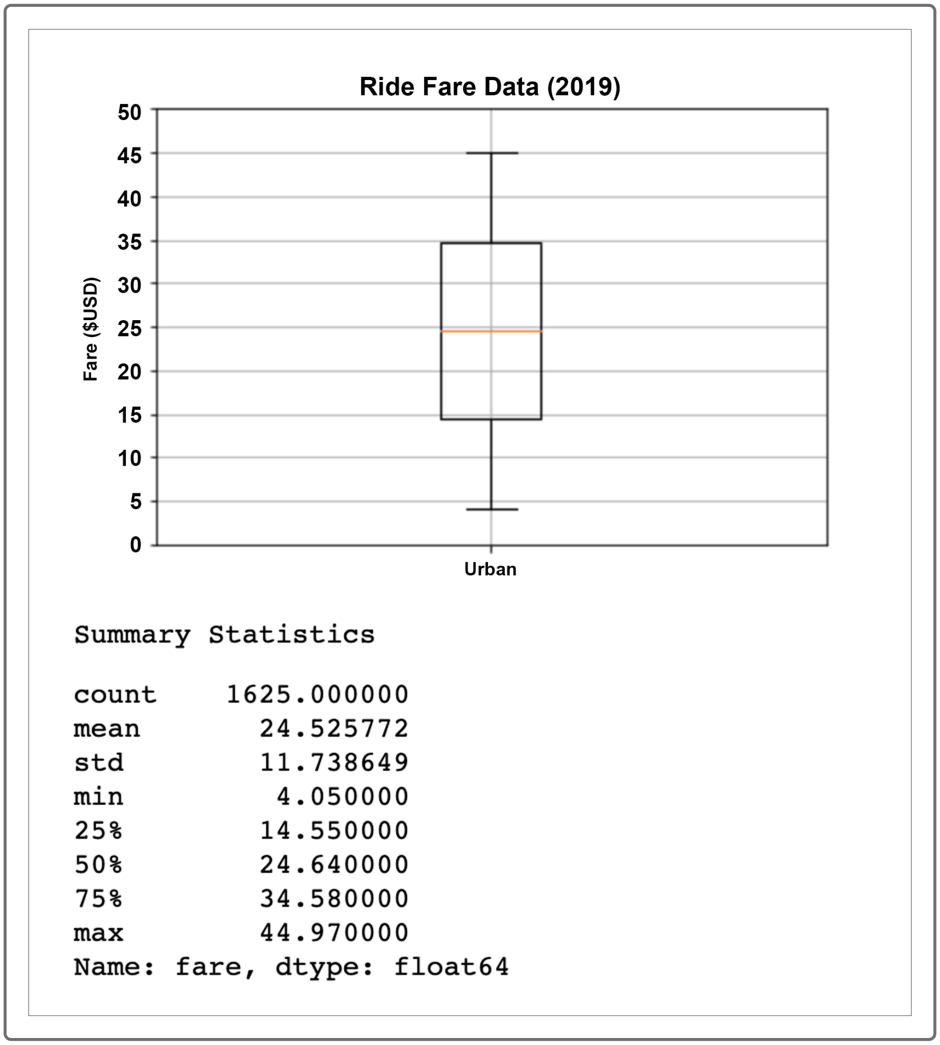
Box-and-Whisker Plots for Ride Fare Data

Next, let's create box-and-whisker plots for the ride fare data with summary statistics.

For the fare data, we will use the `urban_fares` Series we created earlier. Add the following code to the new cell:

```
# Create a box-and-whisker plot for the urban fare data.
x_labels = ["Urban"]
fig, ax = plt.subplots()
ax.boxplot(urban_fares, labels=x_labels)
# Add the title, y-axis label and grid.
ax.set_title('Ride Fare Data (2019)')
ax.set_ylabel('Fare($USD)')
ax.set_yticks(np.arange(0, 51, step=5.0))
ax.grid()
plt.show()
print("Summary Statistics")
urban_fares.describe()
```

When you run the cell, the urban fare data box-and-whisker plot will look like this:



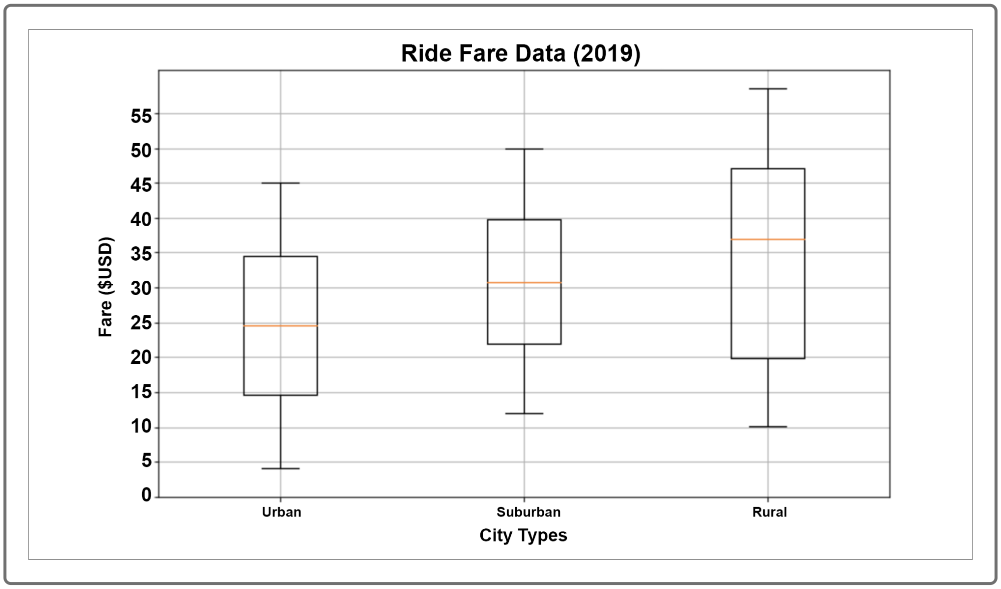
```
Summary Statistics

count    1625.000000
mean       24.525772
std        11.738649
min         4.050000
25%        14.550000
50%        24.640000
75%        34.580000
max        44.970000
Name: fare, dtype: float64
```

SKILL DRILL

Create box-and-whisker plots for the `suburban_fares` and the `rural_fares` with summary statistics.

SKILL DRILL

Create a box-and-whisker plot that has all three city types' fare data in one plot that looks similar to the following image. Save the combined box-and-whisker plot as `Fig3.png` to your "analysis" folder.



FINDING

From the combined box-and-whisker plots, we see that there are no outliers. However, the average fare for rides in the rural cities is about $11 and $5 more per ride than the urban and suburban cities, respectively. Why do you think there is such a big difference? By looking at the number of riders for each city, can you get a sense of the overall revenue?

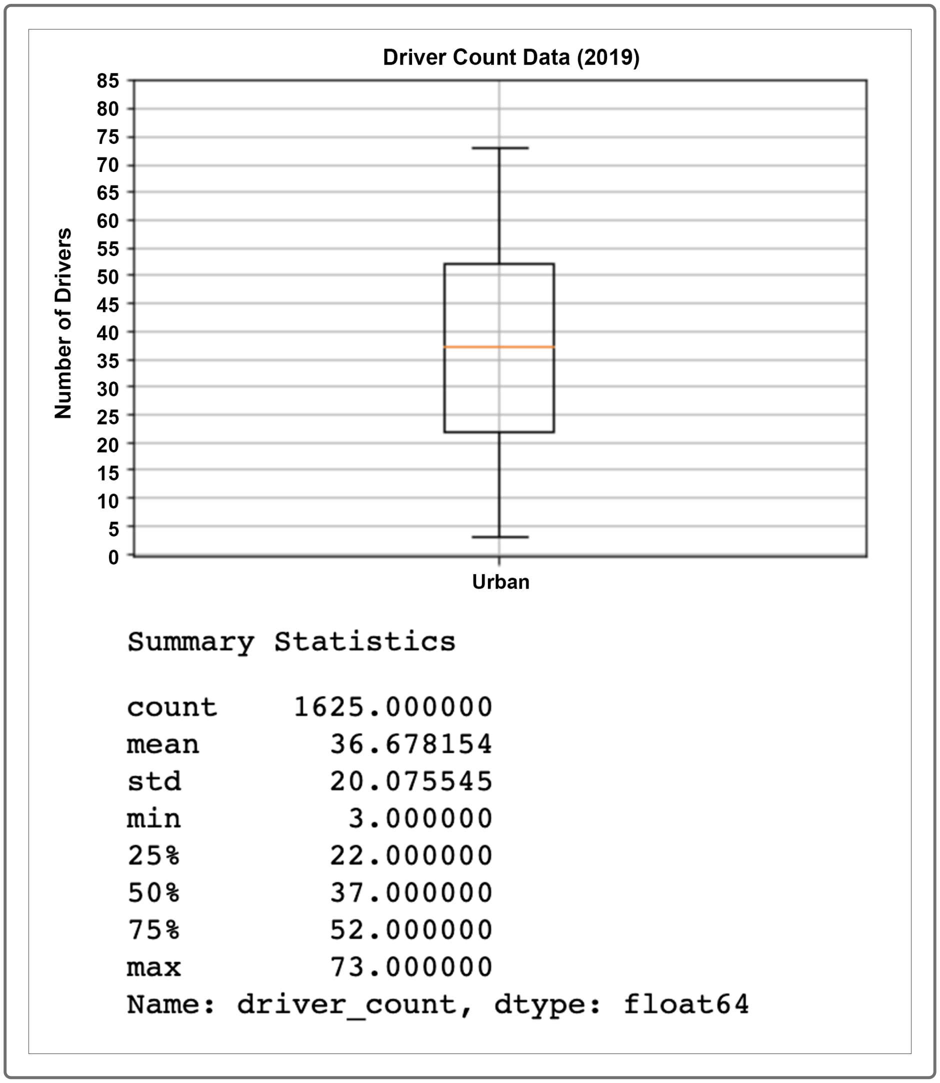
Box-and-Whisker Plots for Driver Count Data

We're getting really good at creating box-and-whisker plots! We need to do one last set of box-and-whisker plots. Let's create a box-and-whisker plot for the driver count data with summary statistics.

For the driver count data, we'll use the `urban_drivers` Series we created earlier. Add the following code to a new cell:

```
# Create the box-and-whisker plot for the urban driver count data.
x_labels = ["Urban"]
fig, ax = plt.subplots()
ax.boxplot(urban_drivers,labels=x_labels)
# Add the title, y-axis label and grid.
ax.set_title('Driver Count Data (2019)')
ax.set_ylabel('Number of Drivers)')
ax.set_yticks(np.arange(0, 90, step=5.0))
ax.grid()
plt.show()
print("Summary Statistics")
urban_drivers.describe()
```

When you run the cell, the urban fare data box-and-whisker plot will look like this:



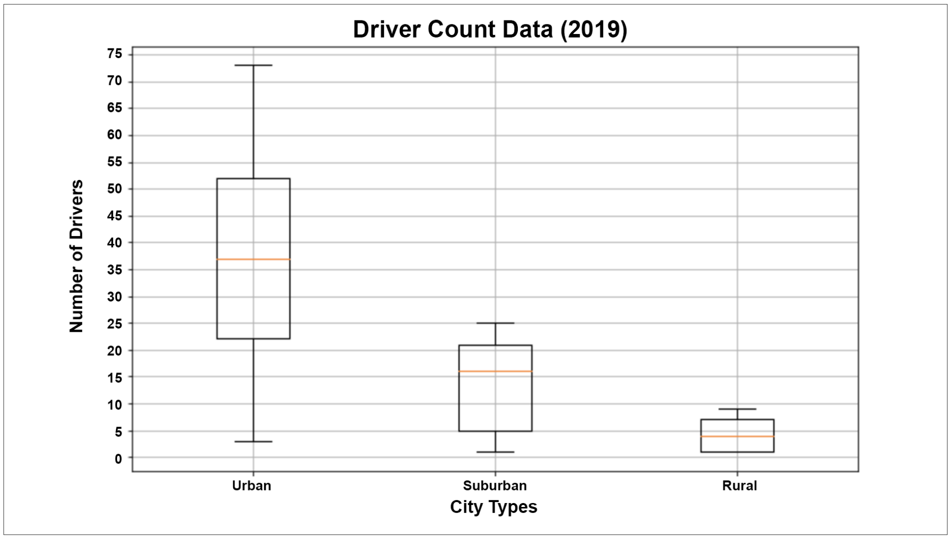
```
Summary Statistics

count    1625.000000
mean       36.678154
std        20.075545
min         3.000000
25%        22.000000
50%        37.000000
75%        52.000000
max        73.000000
Name: driver_count, dtype: float64
```

SKILL DRILL

Using the code for the box-and-whisker plots for the urban drivers, create box-and-whisker plots for the `suburban_drivers` and the `rural_drivers` Series with summary statistics.

SKILL DRILL

Create a box-and-whisker plot that has all three city types' driver count data in one box-and-whisker plot that looks similar to the following image. Save this combined box-and-whisker plot as `Fig4.png` in your "analysis" folder.



FINDING

The average number of drivers in rural cities is nine to four times less per city than in urban and suburban cities, respectively. By looking at the driver count data and fare data, can you get a sense of the overall revenue?

NOTE

> For more information on creating box-and-whisker plots using the object-oriented interface method, see the following documentation:
>
> [Matplotlib documentation on statistics visualizations (https://matplotlib.org/stable/gallery/index.html#statistics)](https://matplotlib.org/stable/gallery/index.html#statistics)
>
> [Matplotlib documentation on box plots (https://matplotlib.org/examples/statistics/boxplot_demo.html)](https://matplotlib.org/examples/statistics/boxplot_demo.html)