1 newpage

2 Data Science: Bridging Principles and Practice

2.1 Part 4: Categorizing Rocket Fuel Data [SOLUTIONS]

2.2 4a. Groupby

EXERCISE: Use `group` to group the handbags by `"rating"`. Aggregate them using the `count` method.

```
In [5]: # group by rating
        rating_counts = handbags.groupby("rating").count()
        rating_counts

Out[5]:         color  price
        rating
        3           1      1
        4           3      3
        5           1      1
```

EXERCISE: We want to look at how conversion rates are different for different days of the week in the Rocket Fuel case. Fill in the ellipses below to group the `ads` data by the day on which a user saw the most ads, then add `mean` as the aggregation function.

Note: there's an extra argument in `groupby` called `as_index` which determines whether or not the groups will be the new DataFrame index. In this case, we don't want the groups as the index, so leave it set to `False`.

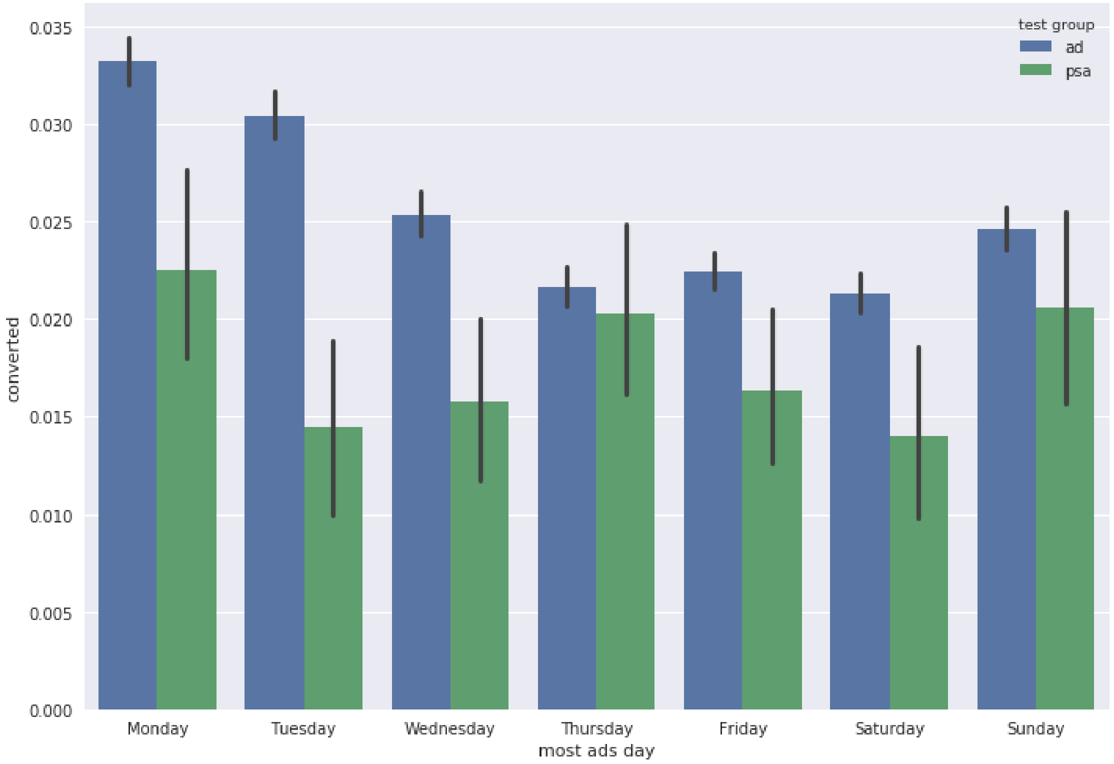
```
In [6]: # fill in the ... with the correct code
        day_rates = ads.groupby("most ads day", as_index=False).mean()
        day_rates

Out[6]:   most ads day       user id  converted  total ads  most ads hour
        0       Monday  1.318831e+06   0.032812  25.328517      14.608179
        1      Tuesday  1.321255e+06   0.029840  23.925464      14.038191
        ... Omitting 1 lines ...
        4       Friday  1.310578e+06   0.022212  26.612129      14.694454
        5     Saturday  1.296354e+06   0.021051  25.227663      14.699963
        6       Sunday  1.300229e+06   0.024476  24.403661      14.384197
```

2.3 4c. Consumer Response vs. Day of Week

```
In [8]: sns.set(rc={'figure.figsize':(11.7,8.27)})

        # make a bar plot with different colors for each test group
        sns.barplot(x="most ads day", y="converted", hue="test group", data=ads);
```



QUESTION: On which days is advertising the most effective? When is it least effective?

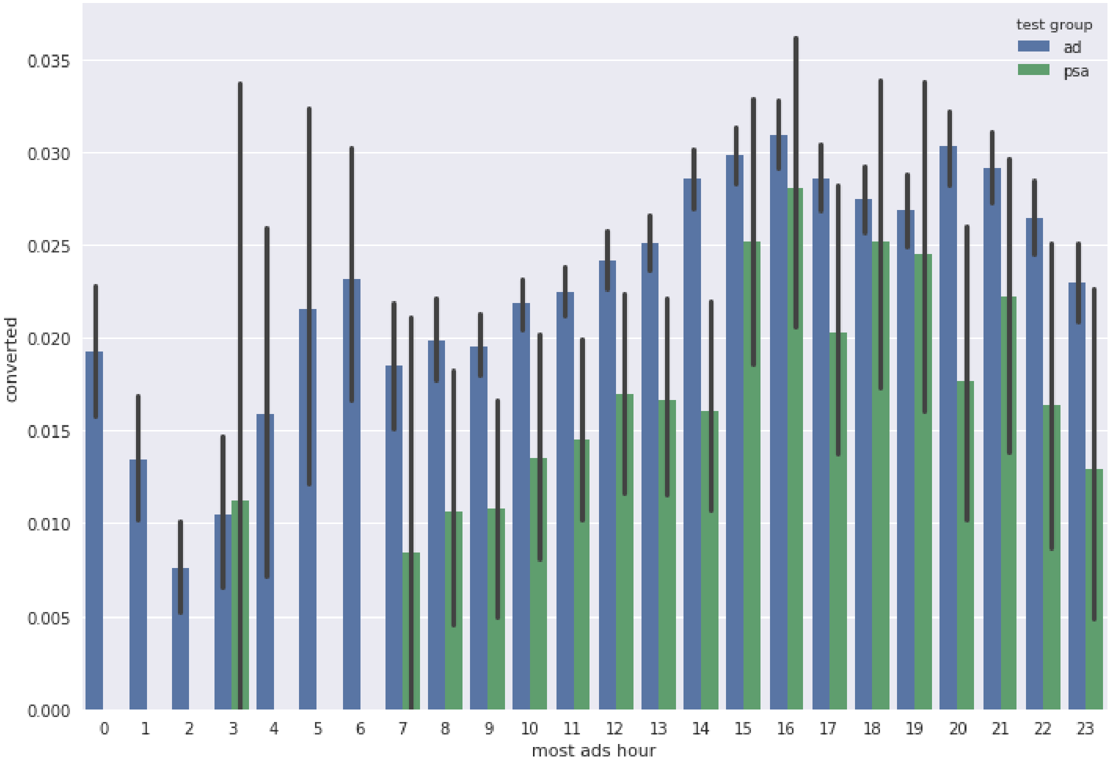
ANSWER: We can think of the efficacy of advertising as the difference between the conversion rate (the rate at which people who saw an image ended up buying a handbag) for the ad group and the psa group. If the coversion rate is *higher* for the ad group than the psa group on a particular day, and all other factors are equal, then we may have evidence that the ad leads more people to buy the bag. If the rate is *equal* for the ad and psa groups, then we're more likely to think the ad is not influencing people's behavior.

The largest difference between the conversion rates for the ad and psa groups occurs on Tuesdays, followed by Mondays. This implies that advertising is most effective early in the week. The smallest difference between the coversion rates, and by implication the least effective advertising day, occurs on Thursdays.

2.4 4d. Consumer Response vs Hour of Day

EXERCISE: Create a bar plot showing the consumer response vs the hour of the day. Look at the "Conversion Rate vs Day of Week and Test Group" example for some hints- you should be able to copy that code and change one value to solve this exercise. Note that when you are referring to the name of a column, you will need quotation marks around the name; when you are referring to the name of a DataFrame, you do not.

```
In [9]: # make a bar plot
        sns.barplot(x="most ads hour", y="converted", hue="test group", data=ads);
```



QUESTION: At which hours is advertising the most effective? When is it least effective?

ANSWER: The largest differences between the ad and PSA bars occur early in the day, between hour 0 (midnight) and hour 6 (6AM) with the greatest difference occurring at hour 6, implying that advertising is most effective during early morning. The smallest differences between the ad and PSA bars occur at hours 16, 18, and 19 (4PM, 6PM, and 7PM respectively), implying that advertising is least effective during the evening.

HOWEVER, you might have been suspicious of the fact that the conversion rates are 0 for so many early morning hours. This is a good example of why it's important to look at the data behind the graph.

If we group subjects by the hour at which they saw the most ads and their test group, then count how many were in each group (code below), we can see that significantly fewer people saw the most ads at an hour between hours 0 and 7 (midnight to 7AM) than any other time of day. For example, there are only 89 people in the hour 3 PSA group (the only time where the conversion rate for the PSA group was actually *higher* than for the ad group) compared to 2,060 in the hour 12 (noon) group. The limited sample sizes for the early hours should lead us to be cautious in how much weight we give to the results.

Hours 8 through 23 had at least 10,000 people in the combined ad and PSA groups. For those hours, we tend to see larger differences between the ad and PSA conversion rates earlier in the morning and later in the evening. The gap between the ad and PSA bars tends to narrow from hours 16 to 19 (4PM to 7PM) and widen as the hours get earlier or later.

```
In [11]: # group the subjects by hour and test group
         ads.groupby(["most ads hour", "test group"]).count()
```

```
Out[11]:                                user id  converted  total ads  most ads day
         most ads hour test group
         0             ad              5309       5309       5309          5309
         ... Omitting 43 lines ...
                       psa              917        917        917           917
```

```
23            ad            19547       19547       19547          19547
              psa             619         619         619            619
```