Project 3

Samantha Uy Tesy

This is the dataset used in this project:

```
# load in data
hotels <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
```

Link to the dataset: https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/readme.md

Part 1

Question: Are hotel booking cancellations more frequent during certain times of the year? If so, do they differ by type of traveler?

Introduction: We are working with the `hotels` dataset, which contains 119,412 rows of hotel booking demand information from 2015 to 2017 for various hotels. Each row represents a single attempted hotel reservation. The dataset contains 32 columns that provide information on the type of hotel, country of origin, date of attempted reservation, reservation dates, customer type information, cancellation information, and average daily rates.

To determine the number of cancellations, we will be working with the following columns:

1. `arrival_date_month`: the month for which the reservation is made for
2. `arrival_date_week`: the week for which the reservation is made for
3. `customer_type`: the type of customer: transient, contract, group, and transient group
4. `is_canceled`: whether the reservation was canceled (1) or not (0)

Approach: Our approach is to first determine if there is exists a difference in cancellations during different months of the year. We will do by creating a pivot table to compare cancellations in each month. Since each month has a varying number of total reservations, we will use a stacked bar chart to compare the proportion of cancellations in each month. Next, we will visualize the number of cancellations for each by customer type to determine if cancellations are driven by a specific type of customer.

To analyze the months with the highest number of cancellations, the following functions will be applied:

1. `group_by()` to group the subsets of interest: month of arrival and whether or not the reservation was canceled
2. `summarize()` to count the number of observations in each group

To plot the proportion of cancellations in each month, the following functions will be applied:

1. `mutate()` to rewrite the `month` column in order
2. `fct_relevel` to manually reorder the `arrival_date_month` column in an intuitive way
3. `geom_bar()` to create a bar plot of the proportion of cancellations

To analyze the weeks with the highest numbers of cancellations, the following functions will be applied:

1. `filter()`: to extract the rows for which a cancellation was made
2. `group_by()`: to group the subsets of interest: customer type
3. `count()`: to count the number of cancellations per week

To plot the the number of cancellations over the year, the following functions will be applied:

1. `geom_line()`: to create a line graph of number of cancellations for each week

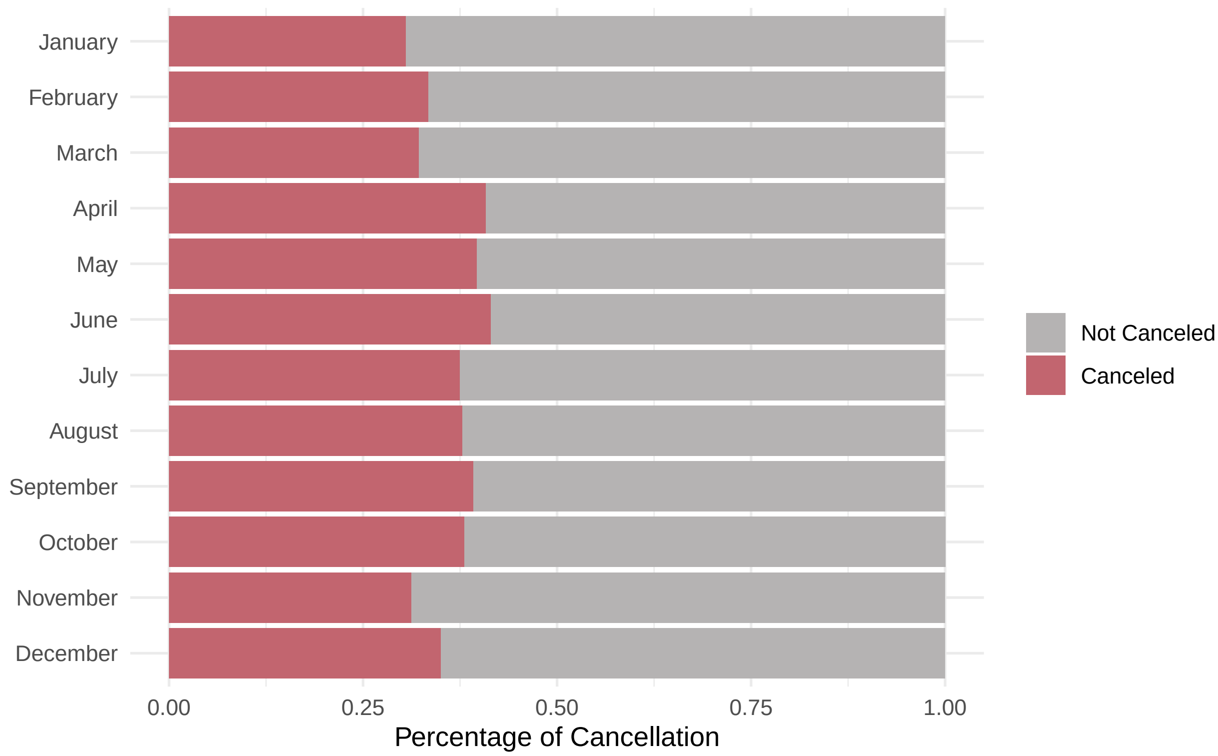
Analysis:

```
# data wrangling
hotels %>%
  group_by(arrival_date_month, is_canceled) %>%
  # return number of observations per group
  summarize(n=n()) %>%

  # create pivot table
  pivot_wider(names_from = "arrival_date_month", values_from = "n")
```

```
## `summarise()` regrouping output by 'arrival_date_month' (override with `.groups` argument)
```

```
## # A tibble: 2 x 13
##   is_canceled April August December February January  July  June March   May
##         <dbl> <int>  <int>    <int>    <int>   <int> <int> <int> <int> <int>
## 1           0  6565   8638     4409     5372    4122  7919  6404  6645  7114
## 2           1  4524   5239     2371     2696    1807  4742  4535  3149  4677
## # ... with 3 more variables: November <int>, October <int>, September <int>
```

```
# data visualization
hotels %>%
  mutate(arrival_date_month = fct_relevel(
    arrival_date_month, "December", "November", "October","September",
                        "August", "July", "June","May", "April",
                        "March", "February", "January")) %>%
  ggplot(aes(y = arrival_date_month, fill = factor(is_canceled))) +
    geom_bar(position = "fill") +
    theme_minimal() +
    scale_x_continuous(
    name = "Percentage of Cancellation"
  ) +
  scale_y_discrete(
    name = NULL
  ) +
  scale_fill_manual(
    # label the legend
    name = NULL,
    labels = c(`1` = "Canceled", `0` = "Not Canceled"),
    # manually set colors of each plot
    values = c(`0` = "#B5B3B3", `1` = "#C2656F"))
```

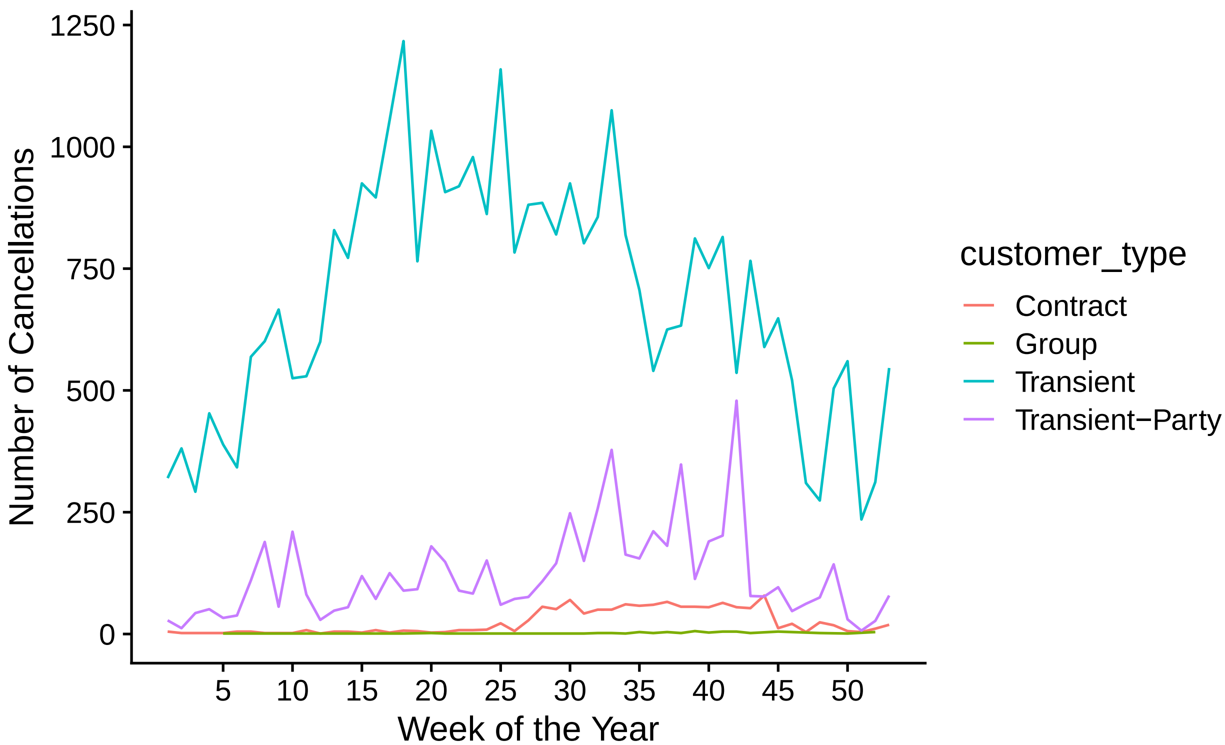



```r
# data wrangling
hotels_week <- hotels %>%
  filter(is_canceled == 1) %>%
  group_by(customer_type) %>%
  count(arrival_date_week_number, is_canceled)

# weekly cancellations
hotels_week
```

```
## # A tibble: 182 x 4
## # Groups:   customer_type [4]
##    customer_type arrival_date_week_number is_canceled     n
##    <chr>                            <dbl>       <dbl> <int>
##  1 Contract                             1           1     5
##  2 Contract                             2           1     2
##  3 Contract                             4           1     2
##  4 Contract                             5           1     2
##  5 Contract                             6           1     5
##  6 Contract                             7           1     5
##  7 Contract                             8           1     2
##  8 Contract                             9           1     2
##  9 Contract                            10           1     2
## 10 Contract                            11           1     8
## # ... with 172 more rows
```

```r
# line graph
hotels_week %>%
  ggplot(aes(x = arrival_date_week_number, y = n, color = customer_type)) +
  geom_line() +
  theme_cowplot() +
```

```r
  scale_x_continuous(
    name = "Week of the Year",
    breaks=c(5, 10, 15, 20, 25, 30, 35, 40, 45, 50)) +
  scale_y_continuous(
    name = "Number of Cancellations") +
    scale_fill_manual(
    # label the legend
    name = NULL)
```



Discussion: Looking at the stacked bar charts, the months with the largest proportion of cancellations are April, June, and September. These months are consistent with popular vacation periods such as Spring break, Summer break, and Labor Day weekend.

Looking at the line graph, it appears the transient and transient-party customer types are more likely to cancel their reservation compared to group and contract types. The highest counts of cancellations for transient type take place at approximately week 17 (April), week 24 (Late May/ June), week 30 (July), and week 33 (September), which is consistent with the conclusions found in the first graph. This may suggest the increased proportion of cancellations are driven a particular group, the transient group types. In the next question, we will explore whether or not families may be behind this trend.

Part 2

Question: Do the average daily rates for families (i.e. customers with children) fluctuate during the year, and are families more likely to cancel during certain times of the year?

Introduction: This question will use the same dataset as Part 1.

To determine how the average daily rates and cancellations for families change over the course of a year, we will be working with the following columns:

1. `arrival_date_month`: the month for which the reservation is made for
2. `adr`: average daily rate (sum of all lodging transactions divided by the number of staying nights)
3. `children`: number of children counted for the reservation
4. `is_canceled`: whether the reservation was canceled (1) or not (0)

Approach: The customer type that has the greatest proportion of children are the transient type. Therefore, our approach is to create a linear model to demonstrate the relationship between average daily rates and the number of children counted for the reservation. This will be used to calculate the mean average daily rates and the standard deviation for the average daily rates. Then we will visualize the distribution over the months of the year using an error bar plot.

To analyze the customer types that have children, these functions will be applied:

1. `group_by()` to group the subsets of interest: customer type, children
2. `summarize()` to count the number of observations in each group

To create a linear model for the distribution of average daily rates:

1. `nest()`: to separate the qualitative data from the linear analysis
2. `mutate()`: to fit a linear model of `adr` on `children`
3. `unnest()`: to add back qualitative columns
4. `select()`: to select the columns of interest
5. `filter()`: filter out "intercept" columns

To plot the distribution of average daily rates over the year, we will use the following functions:

1. `mutate()`: to rewrite the `arrival_month_date` column
2. `fct_relevel()`: to manually reorder the `arrival_date_month` column in an intuitive way
3. `geom_pointrange()`: to create an error bars plot for each month

Analysis:

```
# pivot table to determine which customer types have children
hotels %>%
  group_by(customer_type, children) %>% na.omit() %>%
  # return number of observations per group
  summarize(n=n()) %>%

  # create pivot table
  pivot_wider(names_from = "customer_type", values_from = "n")
```

```
## `summarise()` regrouping output by 'customer_type' (override with `.groups` argument)
```

```
## # A tibble: 5 x 5
##   children Contract Group Transient `Transient-Party`
##      <dbl>    <int> <int>     <int>             <int>
## 1        0     3913   550     81785             24548
## 2        1       68    17      4433               343
## 3        2       93     9      3329               221
## 4        3        1     1        66                 8
## 5       10        1    NA        NA                NA
```

```
# fitting linear model
lm_data <- hotels %>%
  nest(data = -c(arrival_date_month, is_canceled)) %>%
  mutate(
    fit = map(data, ~lm(adr ~ children, data = .x)),
    tidy_out = map(fit, tidy)
  ) %>%
  unnest(cols = tidy_out) %>%
```

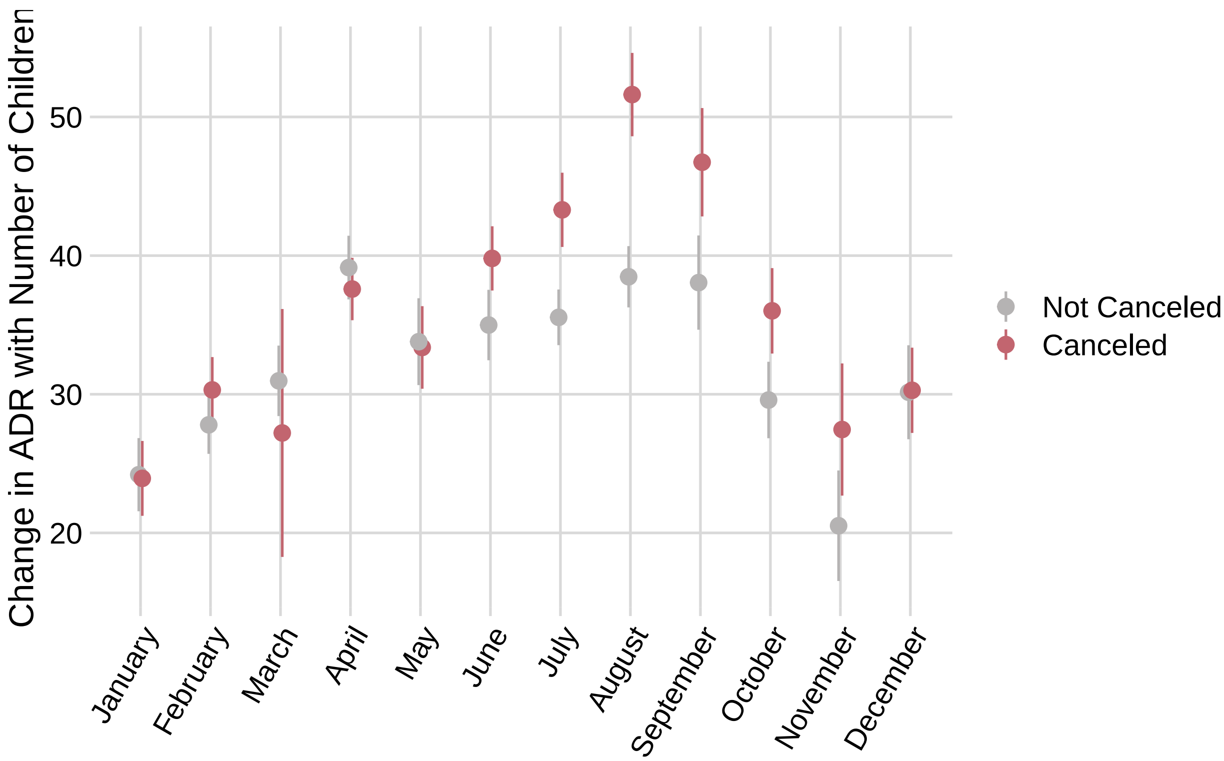
```r
  select(-fit, -data) %>%
  # filter out unnecessary row
  filter(term != "(Intercept)")

# linear model tibble
lm_data
```

```
## # A tibble: 24 x 7
##    is_canceled arrival_date_month term    estimate std.error statistic   p.value
##          <dbl> <chr>              <chr>      <dbl>     <dbl>     <dbl>     <dbl>
##  1           0 July               childr~     35.5      1.02      34.7 4.54e-246
##  2           1 July               childr~     43.3      1.37      31.7 3.12e-200
##  3           0 August             childr~     38.5      1.13      34.1 8.10e-240
##  4           1 August             childr~     51.6      1.53      33.7 6.04e-225
##  5           1 September          childr~     46.7      1.99      23.4 3.72e-114
##  6           0 September          childr~     38.1      1.73      21.9 6.87e-103
##  7           0 October            childr~     29.6      1.41      21.0 3.24e- 95
##  8           1 October            childr~     36.0      1.57      22.9 1.18e-109
##  9           1 November           childr~     27.5      2.43      11.3 9.50e- 29
## 10           0 November           childr~     20.5      2.03      10.1 1.06e- 23
## # ... with 14 more rows
```

```r
# error bars plot
lm_data %>%
  mutate(arrival_date_month = fct_relevel(
    arrival_date_month, "January", "February", "March", "April",
                        "May", "June", "July", "August", "September",
                        "October", "November", "December")) %>%
ggplot(
  aes(
    x = arrival_date_month, y = estimate,
    ymin = estimate - 1.96*std.error,
    ymax = estimate + 1.96*std.error,
    color = factor(is_canceled)
  ))+
  geom_pointrange(
    position = position_dodge(width = 0.1)
  ) +
  scale_x_discrete(
    breaks = unique(hotels$arrival_date_month)
  ) +
  theme_minimal_grid() +
  theme(legend.position = "right",
        axis.text.x = element_text(angle = 60, hjust = 1)) +
  scale_x_discrete(
    name = NULL
  ) +
  scale_y_continuous(
    name = "Change in ADR with Number of Children"
  ) +
  scale_fill_discrete(name = NULL,
                      labels = c(`0` = "A", `1` = "B")) +
  scale_color_manual(
    # label the legend
```

```r
    name = NULL,
    labels = c(`0` = "Not Canceled", `1` = "Canceled"),
    # manually set colors of each plot
    values = c(`0` = "#B5B3B3", `1` = "#C2656F")
  )
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```



Discussion: The months with the highest average daily rate are April, June, July, August, and September. Furthermore, the months in which there is a statistically significant difference between average daily rates (ADR) of reservations that were canceled and not canceled were during the months of June, July, August and September. Additionally, the month of March appears to have the greatest variance of ADR (standard error = 5.96) for canceled reservations, by more than double that of the next highest variance, in November (standard error = 2.43). Conversely, the month with the lowest variance is July (standard error = 1.02).

Looking at the error bars, the ADR seem to increase during the June, July, August, and September months. This is most likely explained by increased demand over the summer break for customers who bring children. Furthermore, the ADR for reservations that were canceled are statistically higher than those that were not canceled. This is intuitive because any customer may be more inclined to cancel a hotel reservation that is too expensive. Thus, we can conclude that some of the increase in cancellations may be driven by customers bringing children.