Telehealth Disparities in Cancer Care Analysis

Tommy Angel

05/21/2024

Note: If you try to Knit this document at this time, you *will* get an error because there is code in this document that has to be edited (by you!) before it will be able to successfully knit!

The Data

Electronic heath record (EHR) data is an electronic version of patients' medical charts. It may include data on a person's demographics, health status, medical care, and medical appointments. One key feature of EHR is that information can be created, managed, and shared across multiple health care organizations and doctors. It is updated and maintained in real time, making new information available instantaneously. A large number of statistical analyses utilize EHR data to answer questions about patient populations and their health outcomes.

This data has been pulled from the Mount Sinai Health System. It contains information on patients seeking oncology treatment. Each observation represents one patient's visit to the Mount Sinai Hospital. Included in the data are main demographic variables such as patient's age, sex, race, and marital status. The data also includes the patient's reason for their medical visit. Finally, the data contains information on the patient's medical appointment, including the time duration and whether the appointment was conducted in-person or remotely using telehealth. You can find the data dictionary in the same folder as the data.

Accessing and Importing the Data

For this project, you'll get the raw data directly from the source. Download the `.csv` file and upload it to `data/raw_data` folder on RStudio.cloud. To load the data in, **run the code in the `ehr-data` code chunk** to create an object called `ehr`.

Note that this data has been deidentified so that it does not contain direct identifiers of patient data. This is an important step in protecting patients' privacy.

```r
ehr <-read_csv("Data/raw_data/ehr_data.csv")
```

```
## Rows: 476 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): gender, race, marital_status, language, insurance, borough, online_...
## dbl (4): id, age, duration, telehealth
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exploratory Analysis of Data

1. The data dictionary is available as `ehr_data_dictionary.pdf`. From the data dictionary, what is the variable that shows whether a patient has set up their online patient portal?

[type answer here: online_portal]

2. From the data dictionary, the variable that represents the telehealth status of appointments can one of take two values, 0 or 1. Which value does 0 represent?

[type answer here: In Person]

3. Describe this data in your own words. Is it observational or experimental? Why? Is this data cross-sectional or longitudinal? Why? How many observations are there?

[type answer here: This data is observational because it involves the observation of patients' characteristics and behaviors without any manipulation by the researcher. It´s cross-sectional because the data was collected at a single point in time and there are 476 observations in the dataset.]

4. How much time, on average, does a patient spend at their appointment?

[type answer here: 40.69328 minutes]

```
### Add Code Here
mean(ehr$duration)
```

```
## [1] 40.69328
```

5. Create a new variable called `duration_hrs`, which represents the time spent at each appointment in hours instead of minutes. What is the longest time someone has spent at an appointment (in hours)?

[type answer here: 2 hours]

```
# create new variable
ehr <- mutate(ehr, duration_hrs = duration / 60)

# get longest appointment
max(ehr$duration_hrs)
```

```
## [1] 2
```

6. We are going to answer whether in-person or telehealth appointments are longer. Start by grouping appointments by their telehealth status and calculate the average time in-person and telehealth visits last. Use the code chunk telehealth-time-analysis in the .Rmd file. Note that you should replace FUNCTION in order to calculate the average of the variable duration.

Which type of appointment lasts longer?

[type answer here: In Person]

```
### Add Code Here
ehr %>%
  # group by telehealth status
  group_by(telehealth) %>%
  # calculate average time per appointment type
  summarize(avg_duration = mean(duration_hrs, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   telehealth avg_duration
##        <dbl>        <dbl>
## 1          0        0.736
## 2          1        0.550
```

7. Use the table function to explore the timeframe variable, which indicates what time of day the medical appointment occurred. What is the most popular timeframe for an appointment? What is the least popular?

[type answer here: Most Popular is 11am - 3pm and the Least Popular is 3pm - 7pm]

```
### Add Code Here
table(ehr$timeframe)
```

```
##
## 11am-3pm  3pm-7pm 7am-11am
##      255       58      163
```

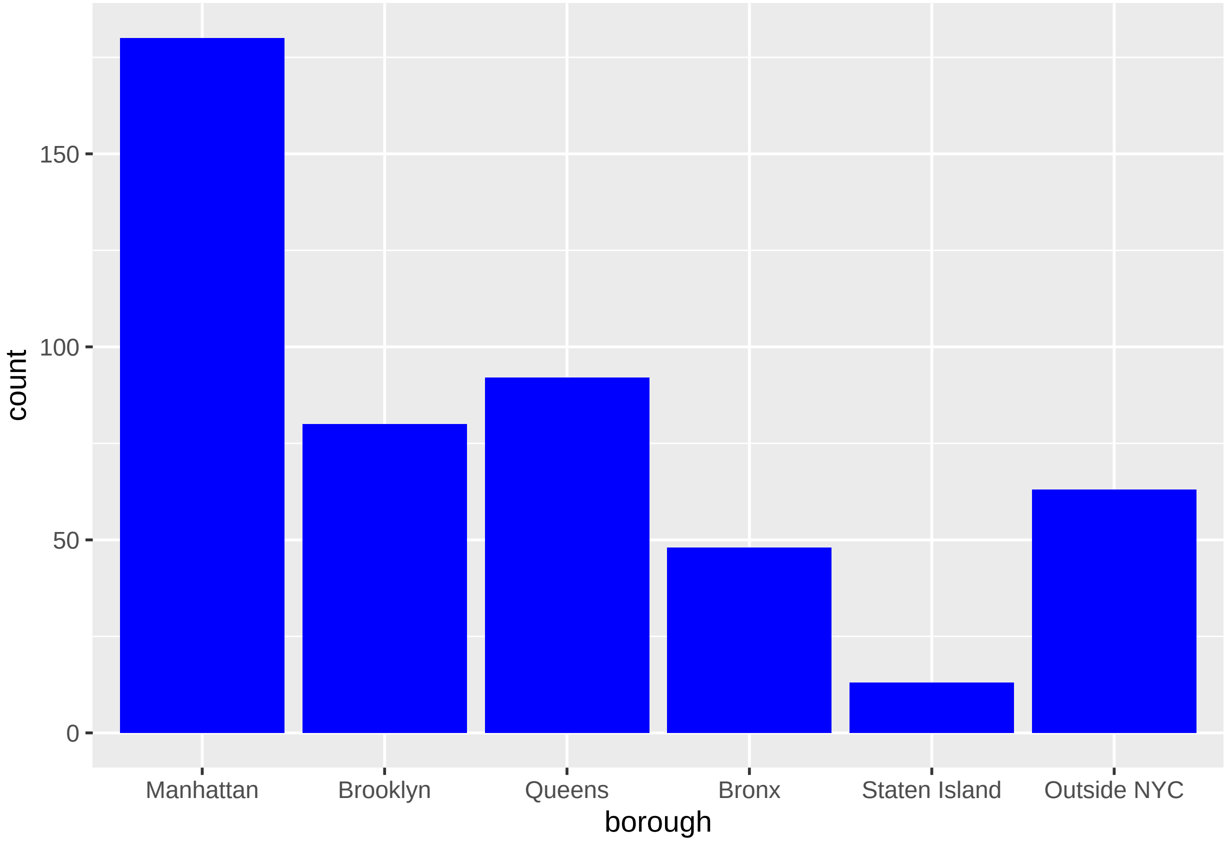
Data Visualization

8. Create a plot that shows the breakdown of how many patients live in each borough. Which borough is most popular? Which is least popular?

[type answer here:Most Popular borough is Manhattan and the Least Popular borough is Staten Island]

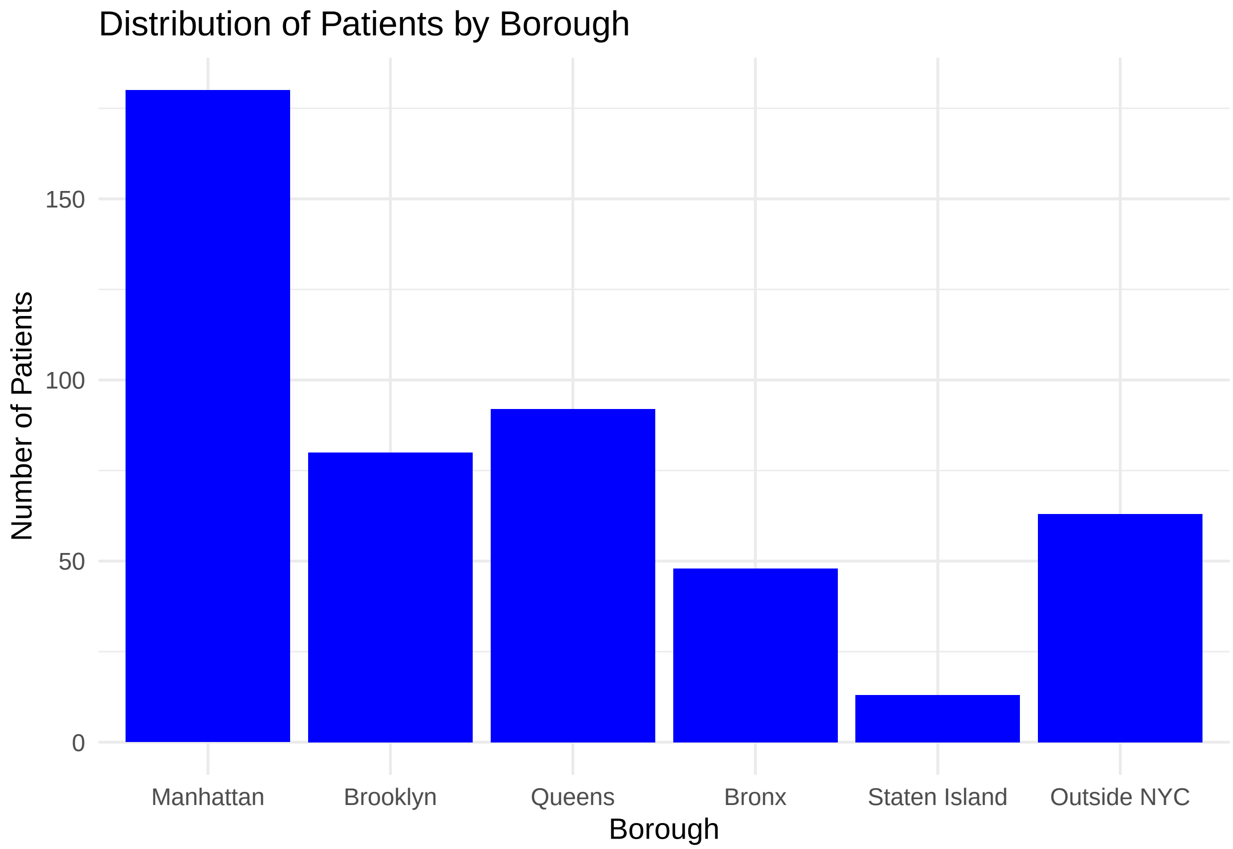
```
# refactor the borough variable
ehr$borough <- fct_relevel(ehr$borough, 'Manhattan', 'Brooklyn', 'Queens', 'Bronx', 'Staten Island', 'Ou

### Add code to generate plot here
ggplot(ehr, aes(x = borough)) +
  geom_bar(fill = "blue")
```



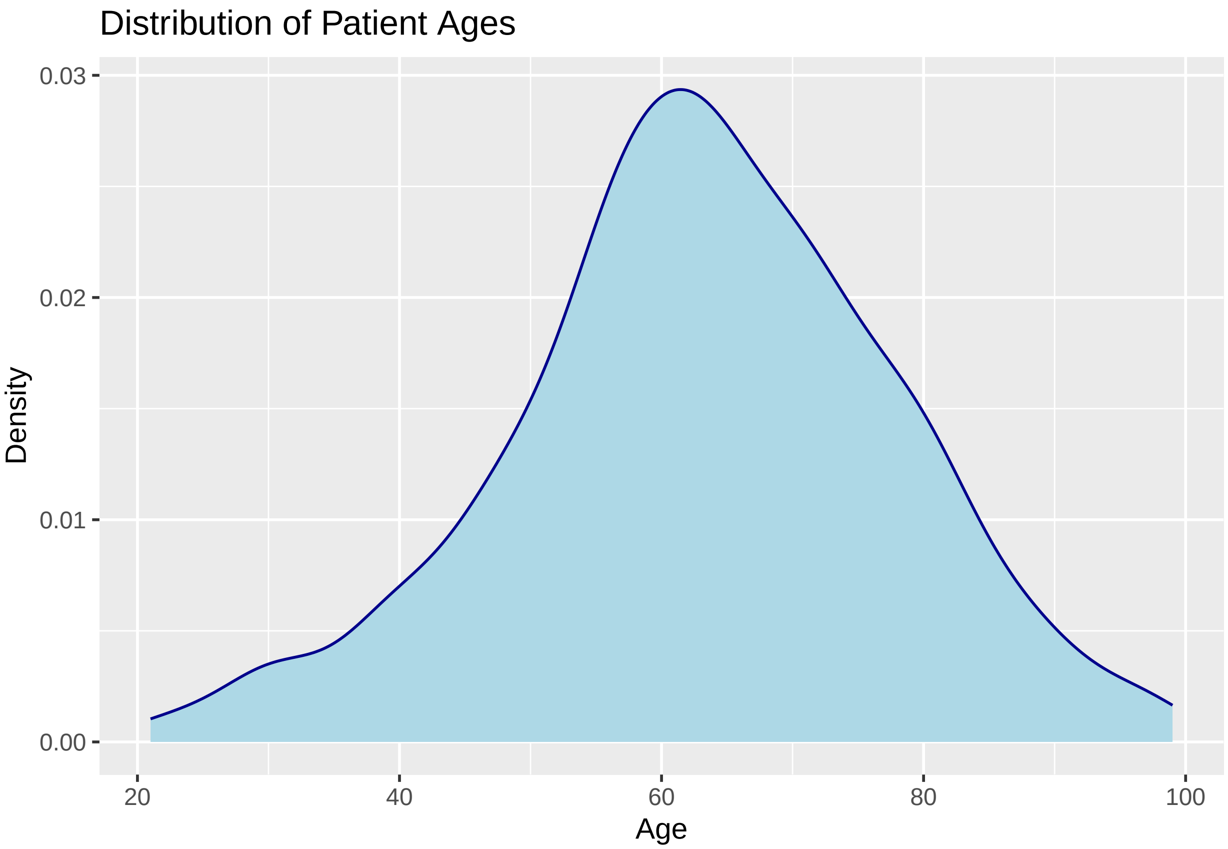
Extend your plot by adding a main title, axis titles, changing the theme, and adding additional features as you prefer. Do the same for all following plots.

```
### Add Code Here
ggplot(ehr, aes(x = borough)) +
  geom_bar(fill = "blue") +
  labs(title = "Distribution of Patients by Borough",
       x = "Borough",
      y = "Number of Patients") +
  theme_minimal()
```



9. Create a density plot of the variable age using ggplot2:

```
### Add Plotting Code Here
ggplot(ehr, aes(x = age)) +
  geom_density(fill = "lightblue", color = "darkblue") +
  labs(title = "Distribution of Patient Ages", x = "Age", y = "Density")
```



Regression Analysis

Now, we are going to explore what factors affect whether a patient chooses to have an in-person or telehealth appointment.

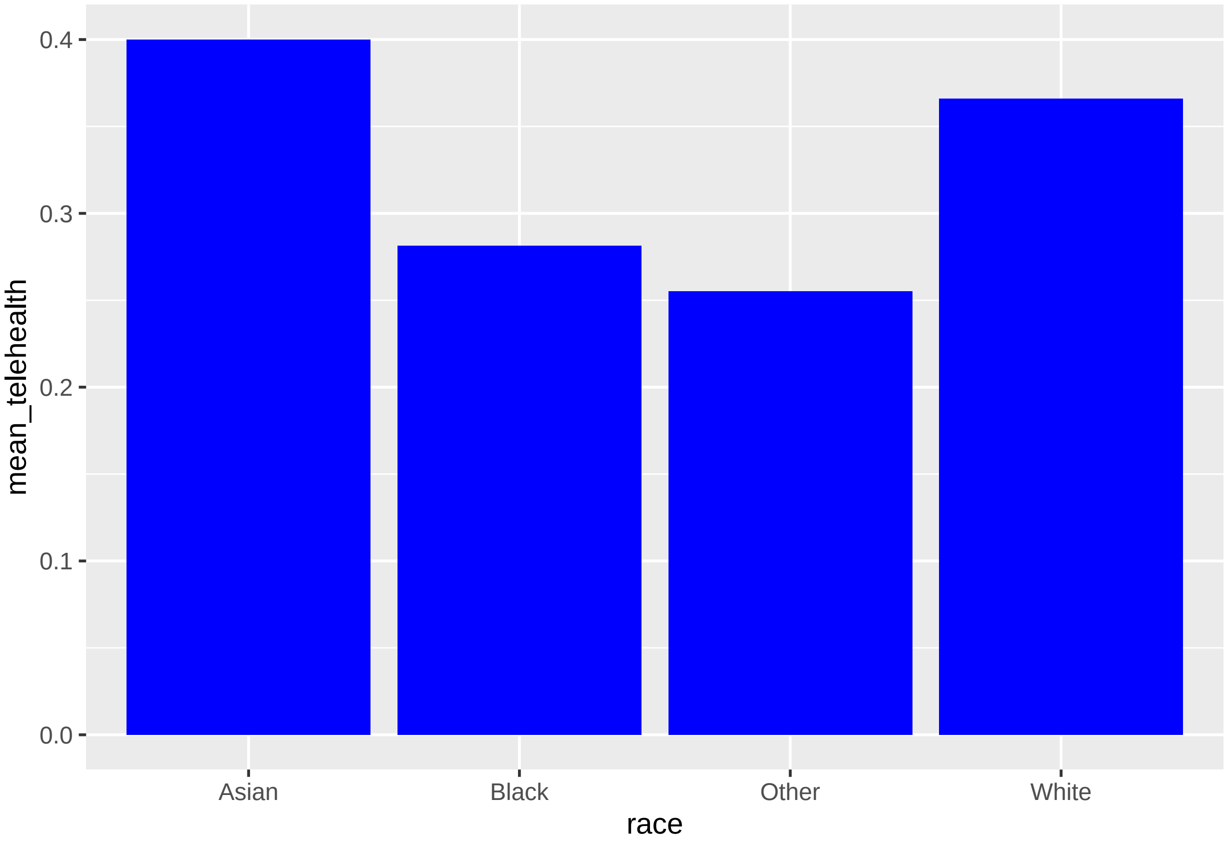
For each of these relationships, present your results in a table or a graph. Feel free to use the data dictionary to find the variables of interest. Add your code in the appropriate `exploratory-analysis` code chunk in the .Rmd file.

The next three questions are fairly open-ended to allow you more freedom in choosing how to explore these relationships and comment on findings.

1. Comment on telehealth usage by race. Use a table or plot to explore this. Remember that it may be more informative to look at proportions instead of raw counts because some groups may naturally be larger in size than others.

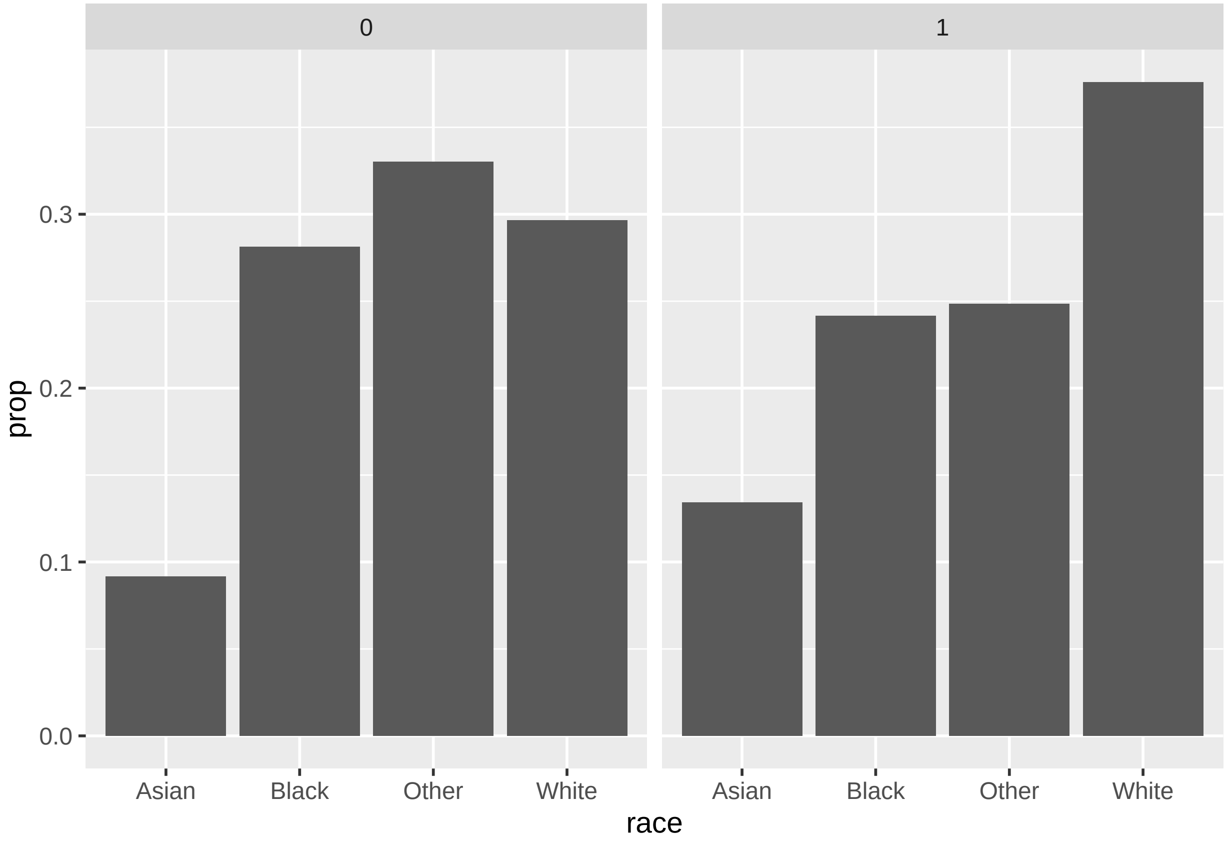
[type your answer here: It seems that there's a higher proportion of Asians that use telehealth than the other races.]

```
## add your exploratory analysis code here
ehr %>%
  group_by(race) %>%
  summarize(mean_telehealth = mean(telehealth == 1)) %>%
  ggplot(aes(x = race, y = mean_telehealth)) +
    geom_bar(stat = "identity", fill = "blue")
```



```
ehr %>%
  group_by(telehealth, race) %>%
  summarize(n = n()) %>%
  mutate(prop = n/sum(n))%>%
  ggplot(aes(x = race, y = prop))+
    geom_bar(stat = "identity") +
    facet_wrap(~telehealth)
```

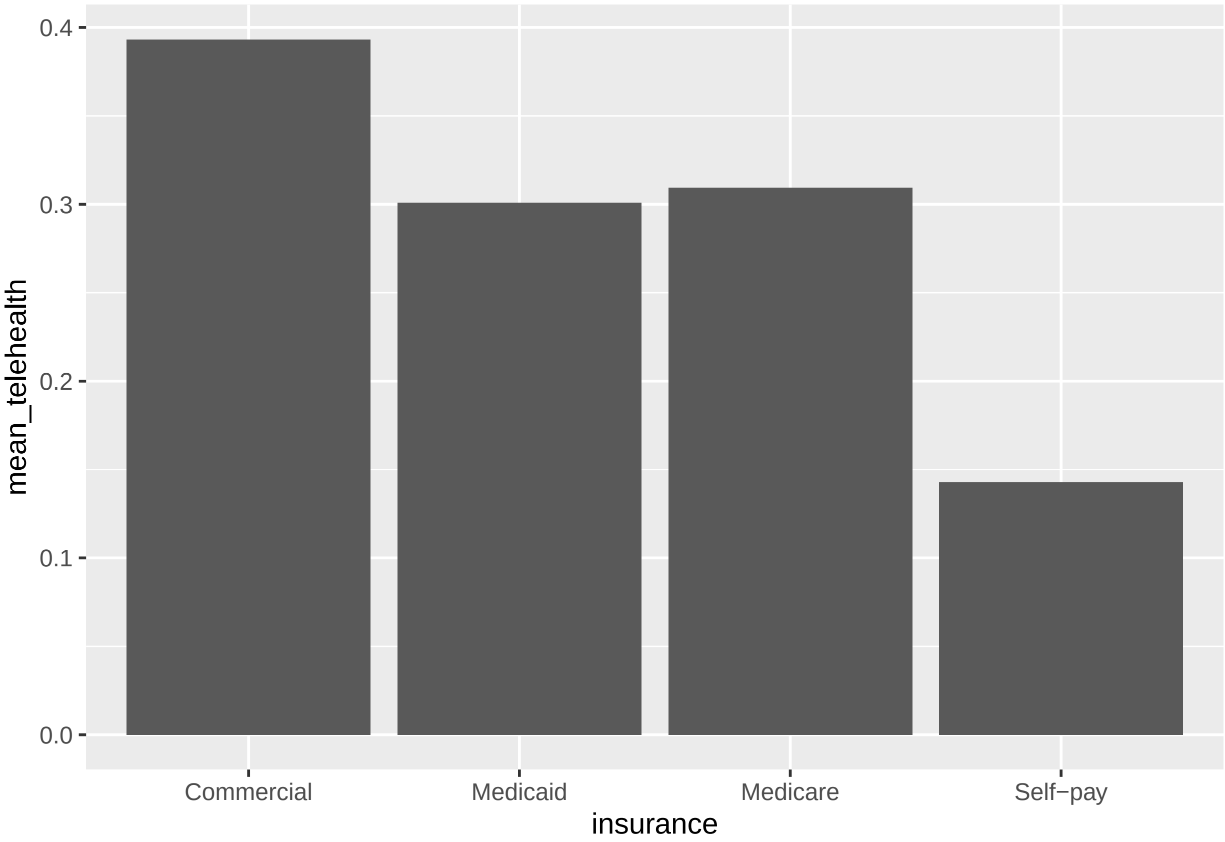
```
## `summarise()` has grouped output by 'telehealth'. You can override using the
## `.groups` argument.
```



2. Comment on the telehealth usage among various insurance groups. Use a table or plot to explore this.

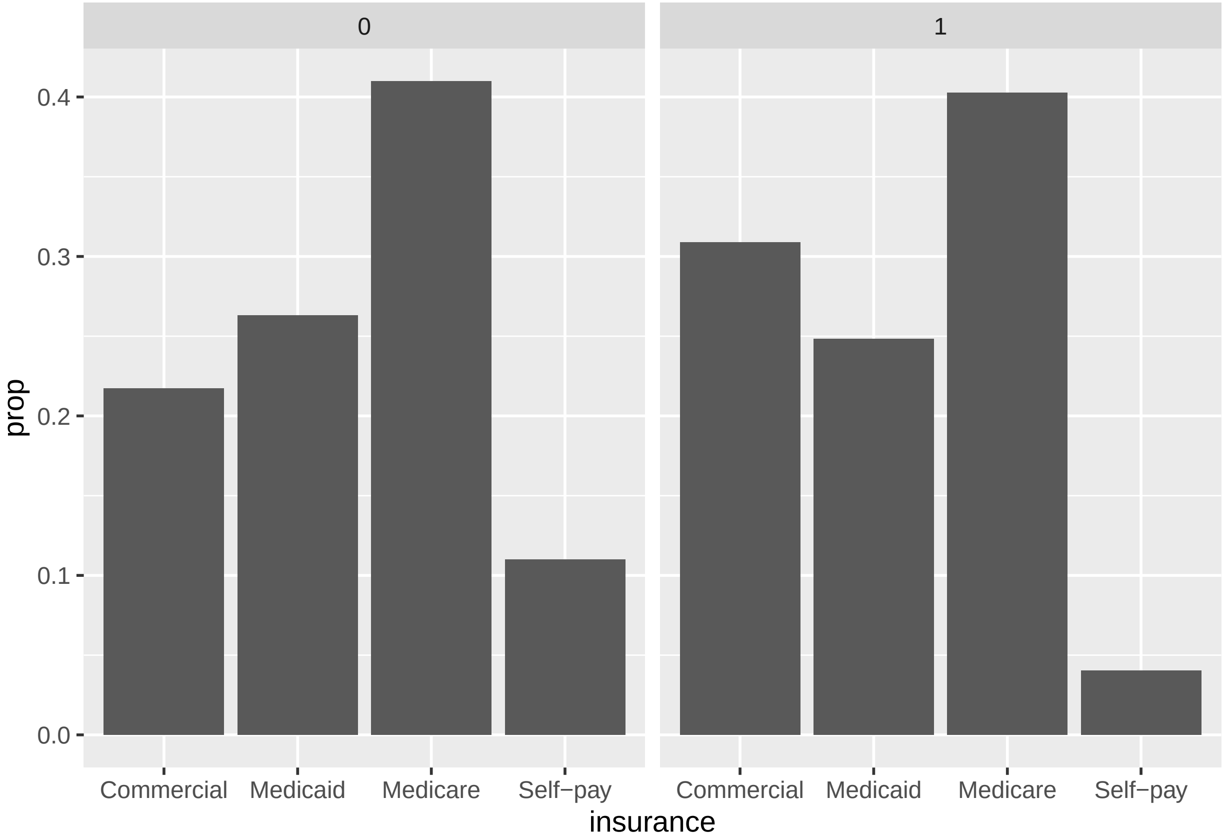
[type your answer here: People with commercial insurance tend to use telehealth as an option more than ones with other insurance types.]

```
## add your exploratory analysis code here
ehr %>%
  group_by(insurance) %>%
  summarize(mean_telehealth = mean(telehealth == 1))%>%
  ggplot(aes(x = insurance, y = mean_telehealth)) +
  geom_bar(stat = "identity")
```



```
ehr %>%
  group_by(telehealth, insurance) %>%
  summarize(n = n()) %>%
  mutate(prop = n/sum(n))%>%
  ggplot(aes(x = insurance, y = prop))+
    geom_bar(stat = "identity") +
    facet_wrap(~telehealth)
```

```
## `summarise()` has grouped output by 'telehealth'. You can override using the
## `.groups` argument.
```



3. Create a new variable called `age_bin` which groups patients into "<65" if they are less than 65 years old or "65+" if they are 65 years or older. Comment on telehealth use by both race and age group. For instance, you can calculate or visualize the difference in proportion of patients by both race and age group for telehealth vs in-person services. Comment on your findings.

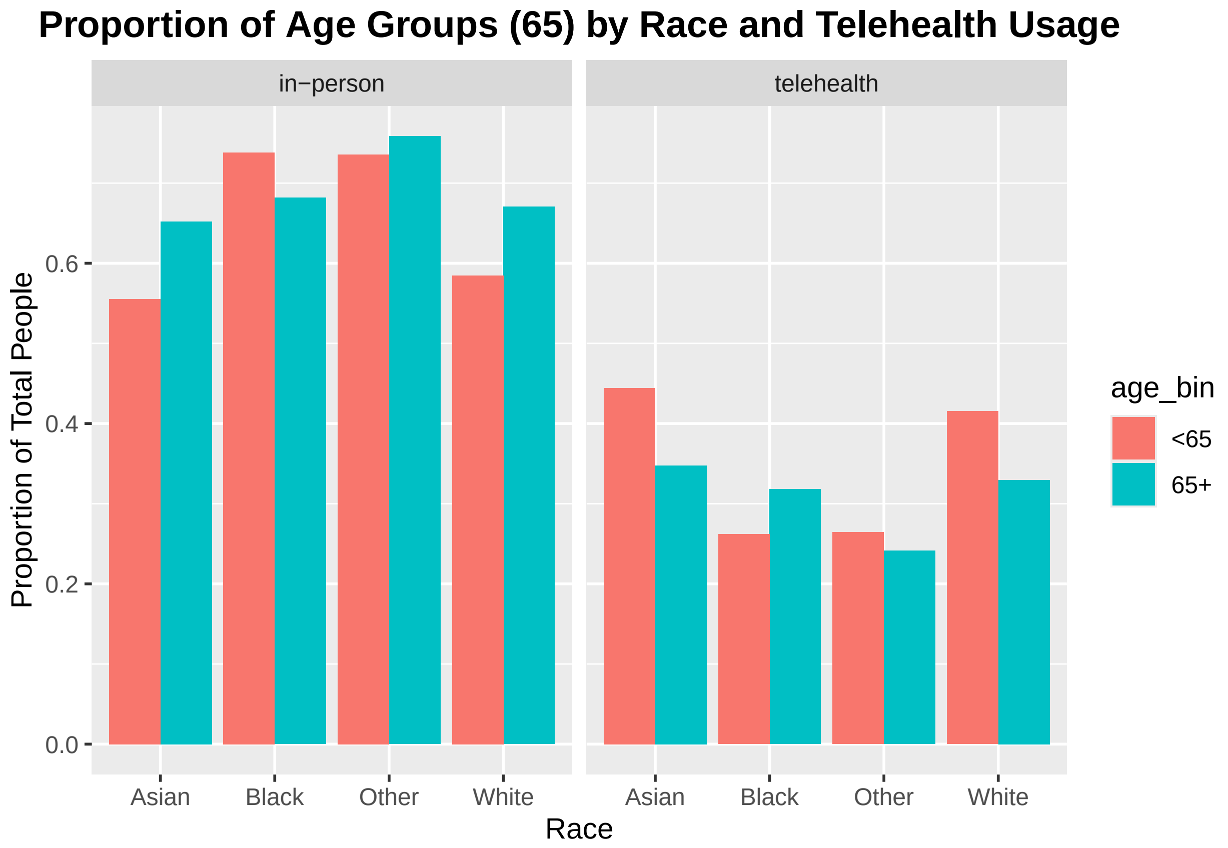
[type your answer here: Regardless of the race, people under the age of 65 still preferred to be in-person rather than use telehealth.]

```r
## create a new variable called age_bin
ehr <- ehr %>%
  mutate(age_bin = ifelse(age < 65, "<65", "65+"))

ehr <- ehr %>%
  mutate(telehealth_bin = ifelse(telehealth == 0, "in-person", "telehealth"),
         telehealth_bin = factor(telehealth_bin))

ehr %>%
  group_by(race, age_bin, telehealth_bin) %>%
  summarize(n = n()) %>%
  mutate(prop = n/sum(n)) %>%
  ggplot(aes(x = race, y = prop, fill = age_bin)) +
    geom_bar(stat = "identity", position = "dodge") +
    facet_wrap(~telehealth_bin) +
    labs(title = "Proportion of Age Groups (65) by Race and Telehealth Usage",
         x = "Race",
         y = "Proportion of Total People") +
    theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
```

```
## `summarise()` has grouped output by 'race', 'age_bin'. You can override using
## the `.groups` argument.
```



4. In the exercise above, we looked at bilateral (two-way) relationships. For instance, we looked at how appointment type and race are related. You have learned in this course, however, that other confounding variables can be a source of bias in your analysis. For instance, the effect of online portal status on appointment type can be biased by the age of the person. Maybe younger people are more likely to have a portal set up because they spend more time online. It's much better to look at the relationship of all relevant variables associated with appointment type together. Run a logistic regression on telehealth status versus race, age, gender, language, insurance, and online portal activation status. Set the reference group for race to be White. Be sure to use the continuous variable for age.

Add your code in the reg-analysis code chunk.

```
ehr$race <- fct_relevel(ehr$race, "White", "Black", "Asian", "Other")
reg_model <- glm(telehealth_bin ~ race + age + gender + language + insurance + online_portal, data=ehr,
summary(reg_model)
```

```
##
## Call:
## glm(formula = telehealth_bin ~ race + age + gender + language +
##     insurance + online_portal, family = "binomial", data = ehr)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.014578   0.578884   1.753 0.079663 .
## raceBlack                  -0.526707   0.275257  -1.914 0.055683 .
## raceAsian                   0.159598   0.383801   0.416 0.677531
## raceOther                  -0.443826   0.284582  -1.560 0.118861
## age                        -0.018265   0.008759  -2.085 0.037045 *
## genderMale                 -0.090494   0.212153  -0.427 0.669706
## languageOther              -0.292271   0.477365  -0.612 0.540367
## languageSpanish            -0.722630   0.491544  -1.470 0.141529
## insuranceMedicaid          -0.188538   0.293313  -0.643 0.520363
## insuranceMedicare           0.067976   0.291499   0.233 0.815610
## insuranceSelf-pay          -1.138620   0.496936  -2.291 0.021947 *
## online_portalInactivated -0.370674   0.361734  -1.025 0.305498
## online_portalPending       -1.277933   0.362502  -3.525 0.000423 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 591.67  on 475  degrees of freedom
## Residual deviance: 549.69  on 463  degrees of freedom
## AIC: 575.69
##
## Number of Fisher Scoring iterations: 4
```

a. What is the coefficient on Black race?

[type answer here: -0.526707]

b. When analyzing results from a logistic regression model, converting the coefficients to an odds ratio is easier to interpret. Convert the above coefficient into an odds ratio.

```
exp(coef(reg_model))
```

```
##              (Intercept)                raceBlack                raceAsian
##                2.7581997                0.5905462                1.1730387
##                raceOther                      age               genderMale
##                0.6415773                0.9819011                0.9134795
##            languageOther          languageSpanish        insuranceMedicaid
##                0.7465665                0.4854737                0.8281694
##        insuranceMedicare        insuranceSelf-pay online_portalInactivated
##                1.0703398                0.3202606                0.6902691
##      online_portalPending
##                0.2786126
```

[type answer here: 0.5905462]

c. Odds ratios can be interpreted as "The odds of group X having outcome Y is [BLANK] higher as compared to the reference group." Interpret the odds ratio for Black patients in the context of the problem.

[type answer here: The odds of raceBlack having a telehealth visit is 0.5905462 times lower as compared to patients who are white. So therefore patients who are black are less likely to partake in telehealth visits than patients who are white.]

d. Is the coefficient for Black patients significant at the 5% level? Is it significant at the 10% level?

[type answer here: yes the coefficient for black patients is significant at the 10% level but not the 5% level.]

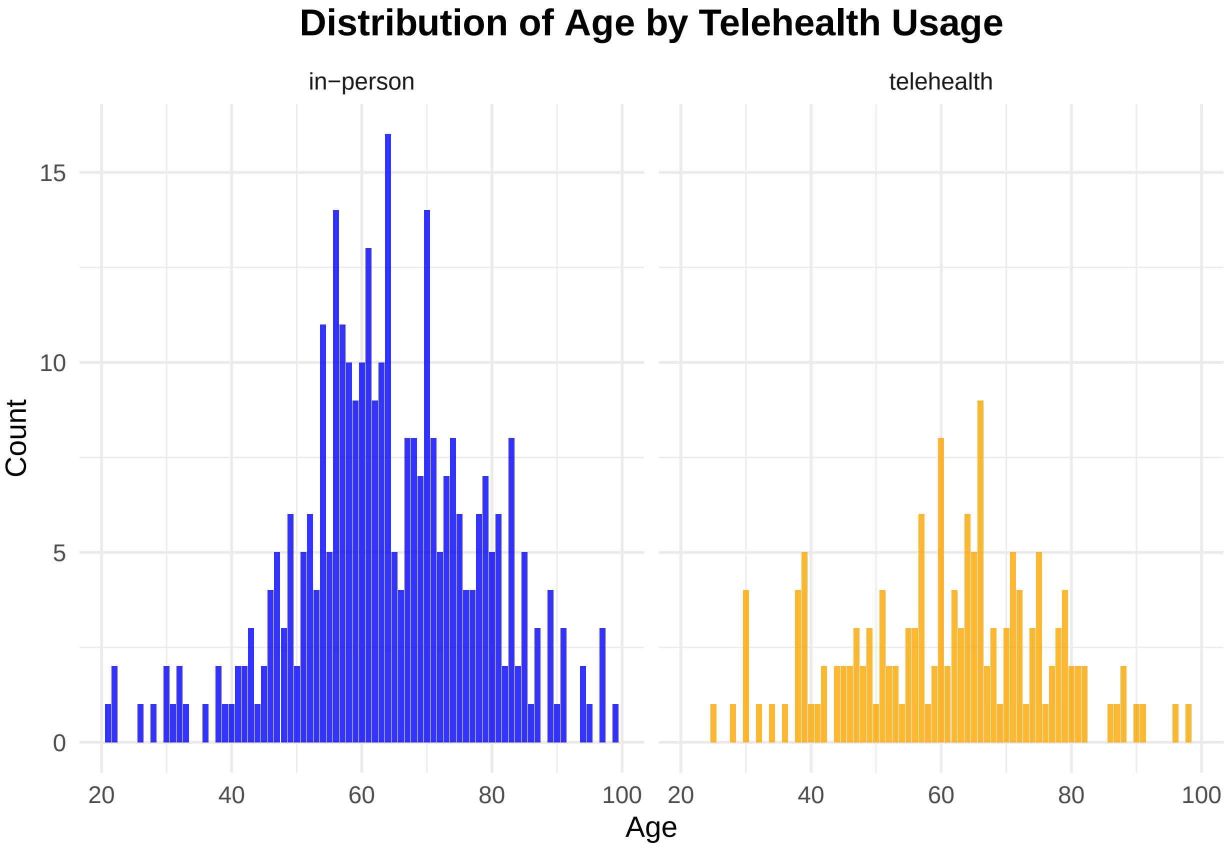
e. As patients get older, are their odds of having a telehealth visit higher or lower than compared to younger patients? Can you think of why this may be the case?

[type answer here: Lower as when age increases the their odds decreases by 0.018265.]

Data Visualization

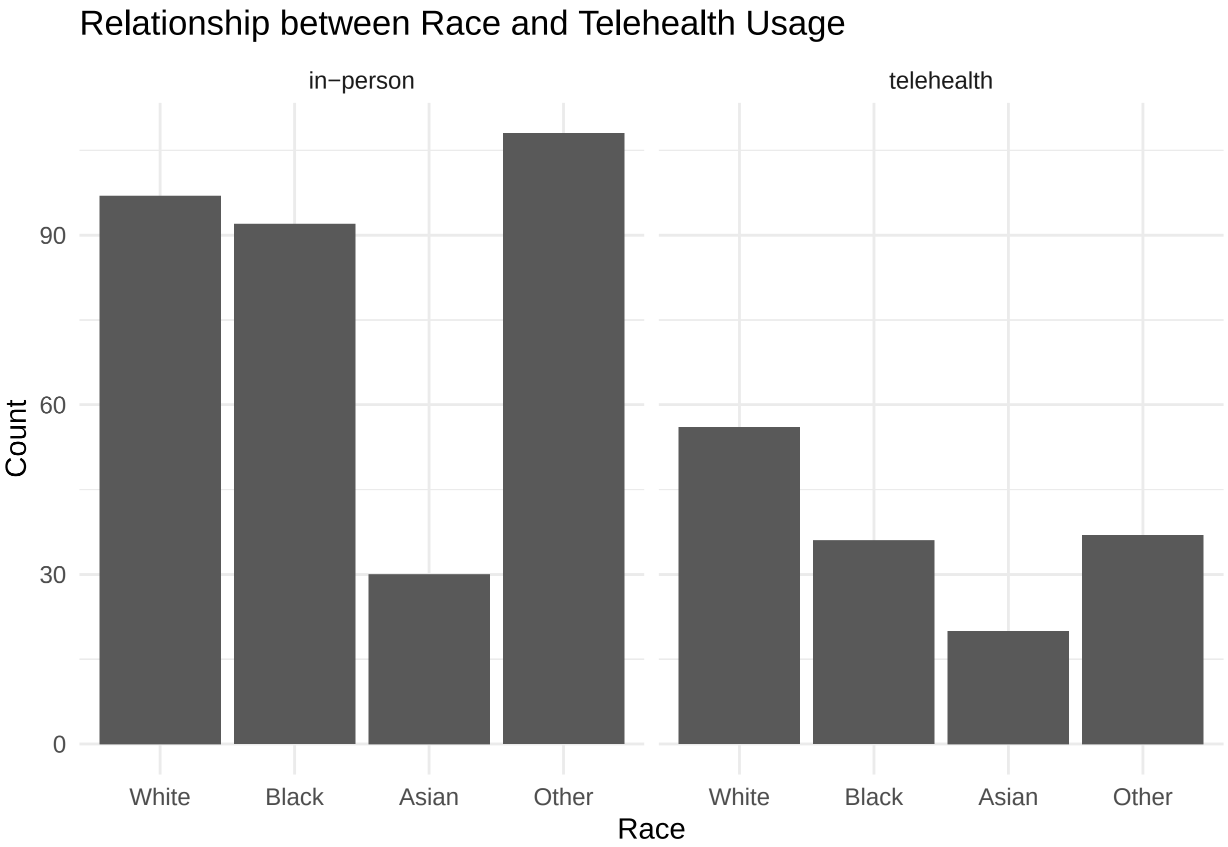
Finally, in this last step, we are going to create a graph that explores the relationship between at least 3 variables at once. Generate a plot that explores the relationship between at least 3 variables (any variable is fine - does not have to be telehealth specific this time). Add any aesthetics or features to make this graph presentable and easy to understand. Remember to add labels and titles. Finally, save the plot to your files.

```r
ggplot(ehr, aes(x = age, fill = telehealth_bin)) +
  geom_bar(position = "dodge", alpha = 0.8) +
  scale_fill_manual(values = c("in-person" = "blue", "telehealth" = "orange")) +
  labs(title = "Distribution of Age by Telehealth Usage",
       x = "Age",
       y = "Count") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"), legend.position = "none"
  ) +
  facet_wrap(~telehealth_bin)
```



```r
ggsave("agetelehealth_plot.png", width = 8, height = 6, units = "in", dpi = 300)
```

```r
ggplot(ehr, aes(x = race)) +
  geom_bar() +
  labs(title = "Relationship between Race and Telehealth Usage",
       x = "Race",
       y = "Count") +
  theme_minimal() +
  facet_wrap(~telehealth_bin)
```



```r
ggsave("TFtelehealth_plot.png", width = 8, height = 6, units = "in", dpi = 300)
```

```r
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 20.04.6 LTS
## 
## Matrix products: default
## BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.8.so;  LAPACK version 3.9.
## 
## locale:
##  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8       
##  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8   
##  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C          
## [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C   
## 
## time zone: UTC
## tzcode source: system (glibc)
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] knitr_1.46      ggthemes_5.1.0  lubridate_1.9.3 forcats_1.0.0  
##  [5] stringr_1.5.1   dplyr_1.1.4     purrr_1.0.2     readr_2.1.5    
##  [9] tidyr_1.3.1     tibble_3.2.1    tidyverse_2.0.0 ggplot2_3.5.1  
## 
```

```
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    stringi_1.8.3     hms_1.1.3
##  [5] digest_0.6.35     magrittr_2.0.3    evaluate_0.23     grid_4.4.0
##  [9] timechange_0.3.0  fastmap_1.1.1     tinytex_0.50      fansi_1.0.6
## [13] scales_1.3.0      textshaping_0.3.7 cli_3.6.2         rlang_1.1.3
## [17] crayon_1.5.2      bit64_4.0.5       munsell_0.5.1     withr_3.0.0
## [21] yaml_2.3.8        tools_4.4.0       parallel_4.4.0    tzdb_0.4.0
## [25] colorspace_2.1-0  vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4
## [29] bit_4.0.5         vroom_1.6.5       ragg_1.3.0        pkgconfig_2.0.3
## [33] pillar_1.9.0      gtable_0.3.5      glue_1.7.0        systemfonts_1.0.6
## [37] xfun_0.43         tidyselect_1.2.1  highr_0.10        rstudioapi_0.16.0
## [41] farver_2.1.1      htmltools_0.5.8.1 rmarkdown_2.26    labeling_0.4.3
## [45] compiler_4.4.0
```

```
ehr$race <- fct_relevel(ehr$race, "White", "Black", "Asian", "Other")
reg_model <- glm(telehealth_bin ~ race + age, data=ehr, family="binomial")
summary(reg_model)
```

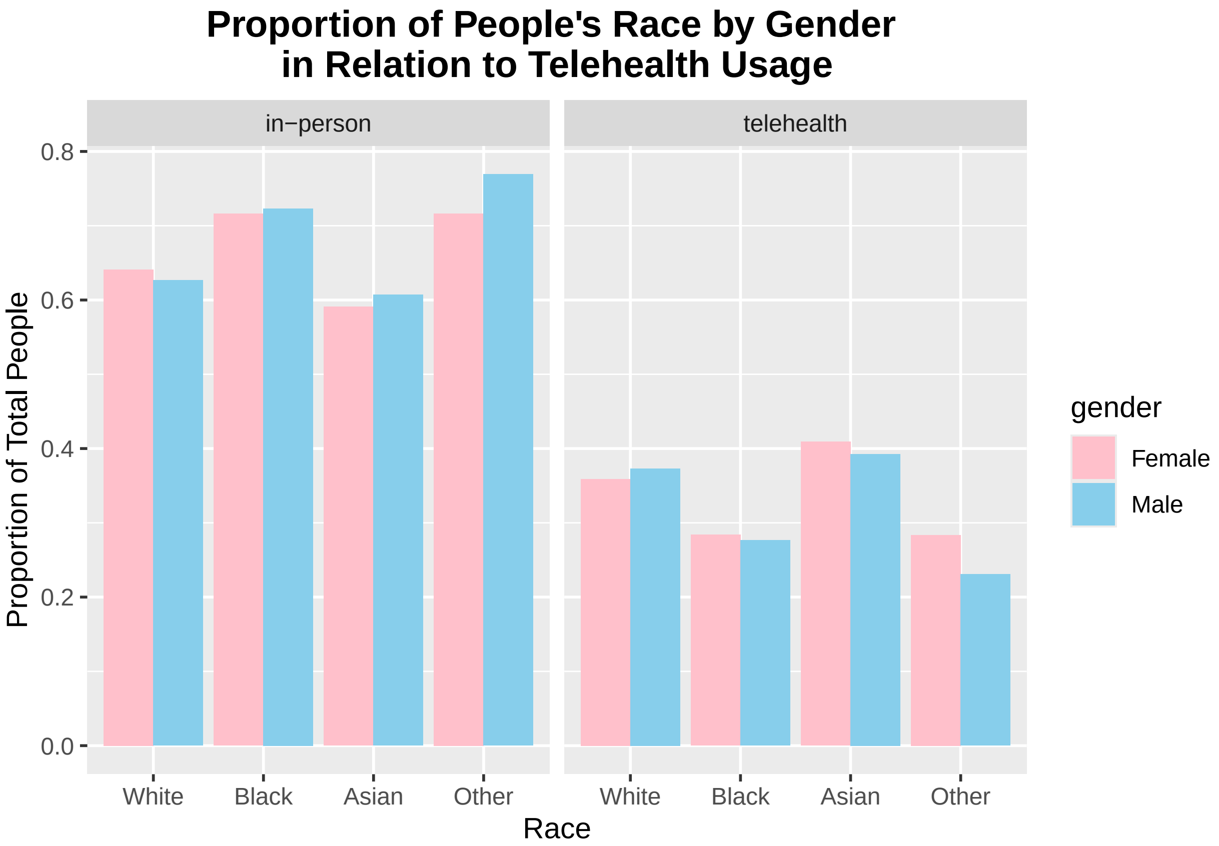
```
##
## Call:
## glm(formula = telehealth_bin ~ race + age, family = "binomial",
##     data = ehr)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.631931   0.490805   1.288   0.1979
## raceBlack   -0.497381   0.264110  -1.883   0.0597 .
## raceAsian    0.057424   0.338400   0.170   0.8653
## raceOther   -0.619479   0.258873  -2.393   0.0167 *
## age         -0.017836   0.007003  -2.547   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 591.67  on 475  degrees of freedom
## Residual deviance: 578.49  on 471  degrees of freedom
## AIC: 588.49
##
## Number of Fisher Scoring iterations: 4
```

```
    exp(coef(reg_model))
```

```
## (Intercept)   raceBlack   raceAsian   raceOther         age
##   1.8812402   0.6081211   1.0591044   0.5382249   0.9823222
```

```
ehr %>%
  group_by(race, gender, telehealth_bin) %>%
  summarize(n = n()) %>%
  mutate(prop = n/sum(n)) %>%
  ggplot(aes(x = race, y = prop, fill = gender)) +
    geom_bar(stat = "identity", position = "dodge") +
    facet_wrap(~telehealth_bin) +
    labs(
      title = "Proportion of People's Race by Gender \nin Relation to Telehealth Usage", x = "Race", y
    scale_fill_manual(values = c("Male" = "skyblue", "Female" = "pink")) +
    theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
```

```
## `summarise()` has grouped output by 'race', 'gender'. You can override using
## the `.groups` argument.
```


Proportion of People's Race by Gender
in Relation to Telehealth Usage
in−person
telehealth
Proportion of Total People
White
Black
Asian
Other
White
Black
Asian
Other
Race
gender
Female
Male