project\_Ta

2024-06-15

library(readxl)  
file\_path <- "/Users/nemotan/Documents/R\_and\_stats/Copy of data\_NT.xlsx"  
project\_data <- read\_excel(file\_path)  
  
str(project\_data)

## tibble [200 × 16] (S3: tbl\_df/tbl/data.frame)  
## $ Sample ID : num [1:200] 1 2 3 4 5 6 7 8 9 10 ...  
## $ Sex : chr [1:200] "F" "F" "F" "F" ...  
## $ MCV : num [1:200] 91.5 86.1 92.7 94.7 94.9 91.7 94.9 92.9 92.5 86.6 ...  
## $ Hemoglobin : num [1:200] 14 14 12 16 14 8 14 12 10 10 ...  
## $ Retic : num [1:200] 1.2 2.6 1.4 2.2 3.9 15.5 1.7 1.9 1.6 1.4 ...  
## $ Platelet count : num [1:200] 274 358 387 338 425 341 414 259 325 50 ...  
## $ Instrument : chr [1:200] "EXN904" "EXN903" "EXN904" "EXN904" ...  
## $ Race : chr [1:200] "Caucasian" "Native American" "Black" "Asian" ...  
## $ Income Bracket : chr [1:200] "High" "Lower-Middle" "Lower-Middle" "Upper-Middle" ...  
## $ Age : num [1:200] 32 40 55 72 59 81 60 52 53 52 ...  
## $ Reason for visit : chr [1:200] "Routine" "Routine" "Obesity" "Injuries" ...  
## $ Season for visit : chr [1:200] "Summer" "Fall" "Winter" "Spring" ...  
## $ Smoke : chr [1:200] "no" "no" "yes" "yes" ...  
## $ Children : chr [1:200] "1" "3" ">3" "1" ...  
## $ Year last seen : num [1:200] 2024 2019 2024 2021 2021 ...  
## $ Exercise regularly: chr [1:200] "yes" "yes" "no" "yes" ...

# Details of data

* Data includes 200 patients’ information at Brigham and Women’s hospital specifically recorded at the Hematology department. Records were selected randomly in the last 5 years when the patients were last seen at the hospital.

# Quantitative variables

* MCV: stands for mean corpuscular volume that measures the average size of red blood cells
* Hemoglobin: is the protein in red blood cells that delivers oxyen to tissue
* Retic: is the immature red blood cells. When released from the bone marrow to early, they need couple days to mature
* Platelet count: is the count of thrombocytes in your blood. They are responsible to reacting to blood vessel injury by clumping and stop bleeding
* Age: age of the patient
* Year last seen: year the patient was last seen at the hospital

# Categorical variables

* Sex: generally only female and male for this data set
* Instrument: The instrument in which the patient’s blood was tested on
* Race: nationality of the patient
* Income Bracket: what patient identifies their income level at
* Reason for visit: reason for visiting the hospital
* Season for visit: what time of the year the visit was made
* Smoke: does the patient smoke?
* Children: R takes it as a categorical variable because of the “>3” but it should be a numerical varible
* Exericise regularly: does the patient exercise regularly?

# Saving new clean project to a new variable CleanData

CleanData <- project\_data

* I will use “CleanData” to manipulate and make graphs. The column Children supposed to be numerical, but due to character “>” and the word “none”, R interprets it as categorical variable. I will mutate the values “>3” to 4 and “none” to 0. Even after that, the column Children still thinks that the column is character variables, not numerical. So, as.numeric will convert the values in Children to numerical.

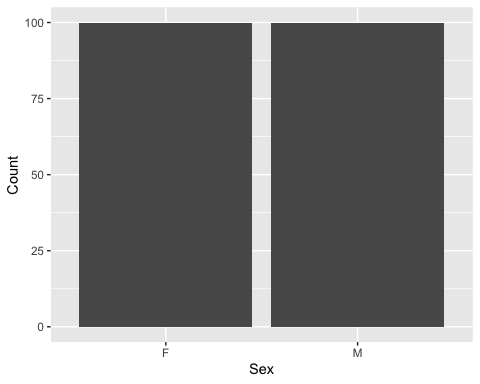
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

CleanData<-  
 CleanData %>%   
 mutate(Children = ifelse(Children==">3", "4", ifelse(Children=="none", 0, Children)))  
CleanData <-  
 CleanData %>%  
 mutate(Children = as.numeric(Children))

# Graph 1 with variable Sex

library(tidyverse)  
ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Sex)) +  
 labs(x = "Sex", y = "Count", color = "engine cylinder")

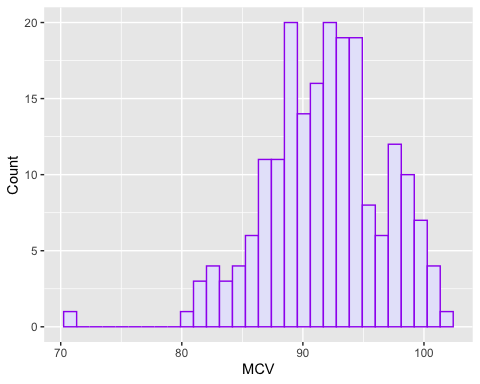


According to this graph, there are an equal amount of sample size between female and male in the data set.

# Graph 2 with variable MCV

ggplot(data = CleanData) +  
 geom\_histogram(mapping = aes(x = MCV), fill = "lavender", color = "purple") +  
 labs(x = "MCV", y = "Count")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



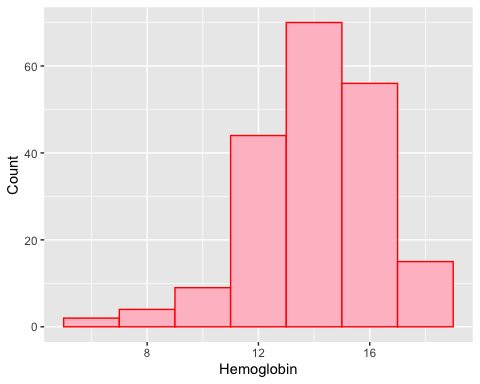
summary(CleanData$MCV)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 71.10 88.60 92.05 91.80 94.90 102.20

The graph above is the histogram that shows the distribution of the Mean Corpuscular Volume (MCV) values. Most of the MCV values are clustered between 80 and 100, with a peak around the 90-95 range. This is a normal MCV range in adults. Although there are few outliers, the values are very few indicating that extreme values are rare. The histogram suggests a somewhat bell-shaped distribution, typical of a normal distribution, but with slight skewness to the left (more frequent lower values).

# Graph 3 with Hemoglobin

ggplot(data = CleanData) +  
 geom\_histogram(mapping = aes(x = Hemoglobin), binwidth = 2, fill = "pink", color = "red") +  
 labs(x = "Hemoglobin", y = "Count")



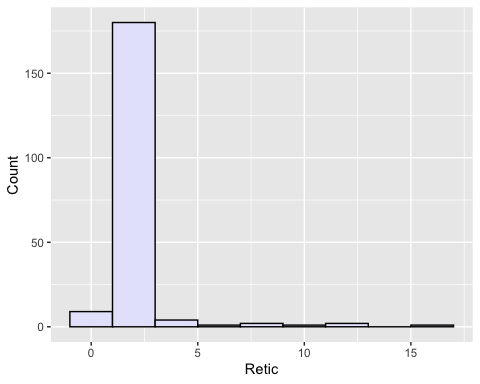
summary(CleanData$Hemoglobin)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.50 13.00 15.00 14.43 16.00 18.00

The histogram above shows the distribution of hemoglobin levels. The majority of the hemoglobin values fall between 12 and 16, with the highest concentration around 14-15. The median (15.00) is higher than the mean (14.43), indicating a slight left skew, with more frequent lower values pulling the mean down. Overall, the hemoglobin levels show a slight skewness to the left with a high concentration around the median value, indicating that while most individuals have typical hemoglobin levels (14-15), there are a few with lower levels that slightly affect the average.

# Graph 4 with variable Retic

ggplot(data = CleanData) +  
 geom\_histogram(mapping = aes(x = Retic), binwidth = 2, fill = "lavender", color = "black") +  
 labs(x = "Retic", y = "Count")



summary(CleanData$Retic)

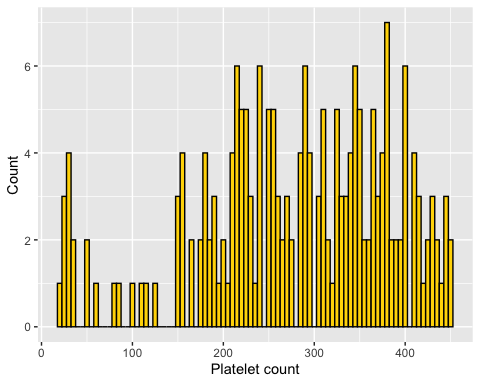
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.700 1.500 1.800 2.096 2.100 15.500

The graph above is a histogram of Retic ranges based on 200 samples. The majority of retic values fall between 0.5 to 2.5 which is the normal range in adults. Retic is a valuable measurement when it comes to blood work that can give health care providers clues about what is going with the patient’s blood or bone marrow. The graph has couple outliers to the right with abnormally high retic value. We shall see in the later graph if there is correlation to why their retic is so high.

# Graph 5 for variable platelet count.

I will rename the column so that it doesn’t have any space in the column name.

CleanData<-  
 CleanData %>%   
 rename(Platelets = `Platelet count`)  
ggplot(data = CleanData) +  
 geom\_histogram(mapping = aes(x = Platelets), binwidth = 5, fill = "gold", color = "black") +  
 labs(x = "Platelet count", y = "Count")



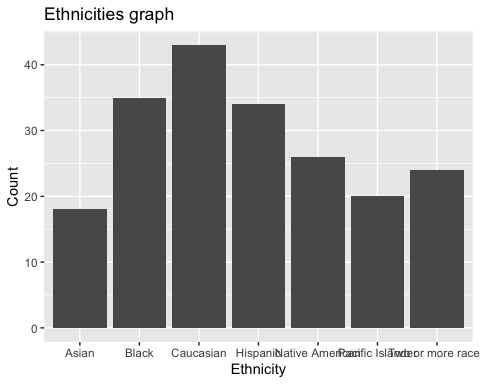
summary(CleanData$Platelets)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.0 215.0 290.0 278.7 363.5 450.0

The histogram above shows a graph of platelet counts of 200 samples. The graph shows majority of values are from 150 to 450 which is normal range of platelet counts in adults. Although it is still slightly left skewed due to some outliers less than 100 platelet count.

# Graph 6 with variable Race

ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Race)) +  
 labs(title = "Ethnicities graph", x = "Ethnicity", y = "Count")

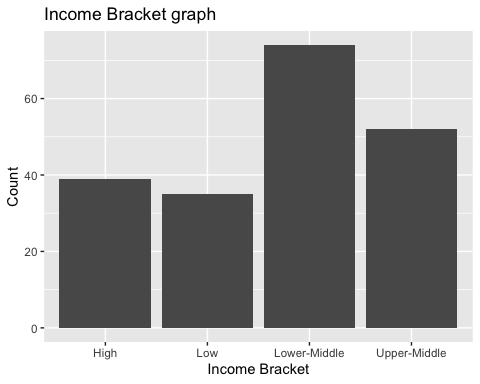


The above bar chart shows the ethnicities of the 200 patients in the dataset. Majority is Caucasion but not too much of a difference between the rest of the group. We could say that among the group, Asian ethnicity tends to not go to the hospital as frequently as the rest based on this chart.

# Graph 7 with variable Income Bracket.

I will rename the column so that it doesn’t have any space in the column name.

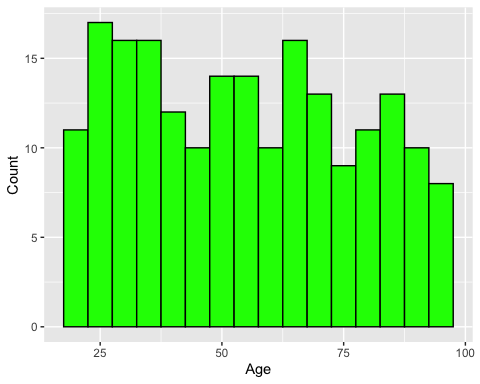
CleanData<-  
 CleanData %>%   
 rename(Income\_Bracket = `Income Bracket`)  
ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Income\_Bracket)) +  
 labs(title = "Income Bracket graph", x = "Income Bracket", y = "Count")



The bar chart aboves shows 4 different income brackets for the 200 samples in the data set. It appears that the majority of participants belong to the lower-middle income bracket. We would have see more covariation graphs to make more sense of the entire dataset whether income bracket affects healthcare or not.

# Graph 8 with variable Age

ggplot(data = CleanData) +  
 geom\_histogram(mapping = aes(x = Age), binwidth = 5, fill = "green", color = "black") +  
 labs(x = "Age", y = "Count")



summary(CleanData$Age)

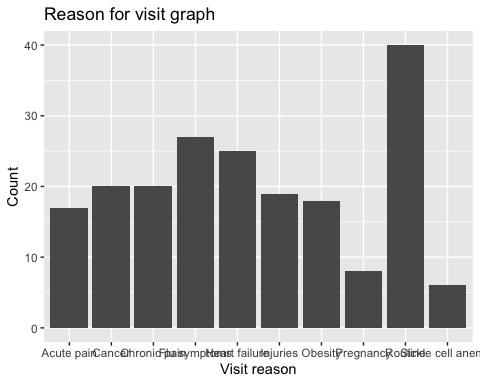
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 34.00 53.00 54.76 73.00 97.00

The histogram above shows the age of 200 samples in the data set. It appears that the dataset includes participants who are adults from age 18 to approximately 98. The distribution is quite normal with the mean almost the same the median. Therefore, the data set has a good range that includes a variety of adult ages.

# Graph 9 Reason for visit.

I will rename the column so that it doesn’t have any space in the column name.

CleanData<-  
 CleanData %>%   
 rename(Visit\_Reason = `Reason for visit`)  
ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Visit\_Reason)) +  
 labs(title = "Reason for visit graph", x = "Visit reason", y = "Count")

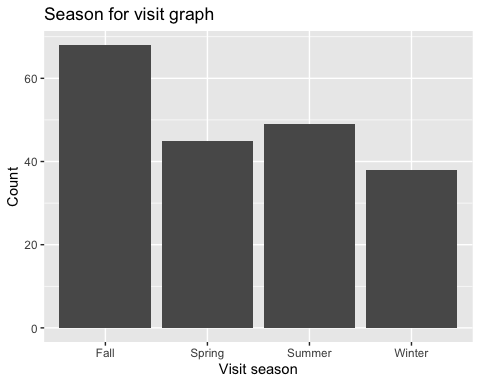


The above bar chart shows a variety of reasons why the 200 patients visited the hospital. It appears that the majority is there for routine check up which is a good thing to do yearly. The second most reason is due to having flu-like symptoms.

# Graph 10 with variable Season for visit.

I will rename the column so that it doesn’t have any space in the column name.

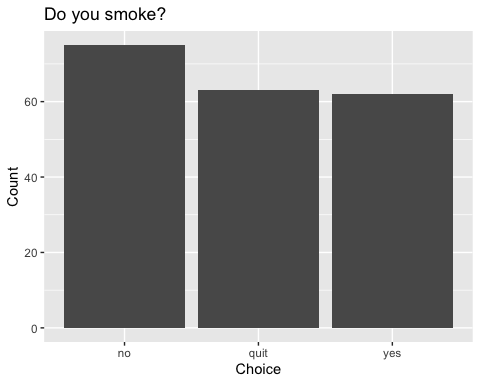
CleanData<-  
 CleanData %>%   
 rename(Visit\_Season = `Season for visit`)  
ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Visit\_Season)) +  
 labs(title = "Season for visit graph", x = "Visit season", y = "Count")



The above graph shows that the people visit the hospital in Fall is slightly higher than the rest of the year. This makes sense since the graph above stated that flu-like symptoms is the second most reason why people visit hospitals. Therefore, the bar where Fall is higher would completely correlate with the reason for visit graph.

# Graph 11 with variable Smoke.

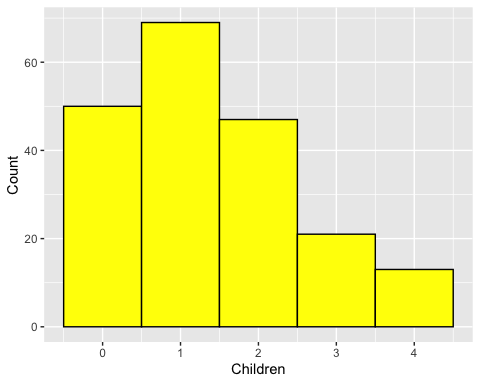
ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Smoke)) +  
 labs(title = "Do you smoke?", x = "Choice", y = "Count")



The above graph shows the choices if the patient smokes or not? This variable would make a valueable fill value for another graph that wants to show if smoking affects your health or not.

# Graph 12 with variable Children

ggplot(data = CleanData) +  
 geom\_histogram(mapping = aes(x = Children), binwidth = 1, fill = "yellow", color = "black") +  
 labs(x = "Children", y = "Count")



summary(CleanData$Children)

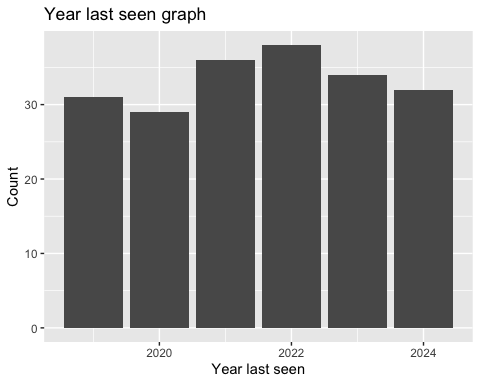
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.75 1.00 1.39 2.00 4.00

The above histogram shows that the majority of the participants in the dataset have 1 child. Then second most common case is that they do not have children or they have 2. The amount of people who have more than 3 children are not that common.

# Graph 13 with variable Year last seen.

I will rename the column so that it doesn’t have any space in the column name.

CleanData<-  
 CleanData %>%   
 rename(Last\_seen = `Year last seen`)  
ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Last\_seen)) +  
 labs(title = "Year last seen graph", x = "Year last seen", y = "Count")

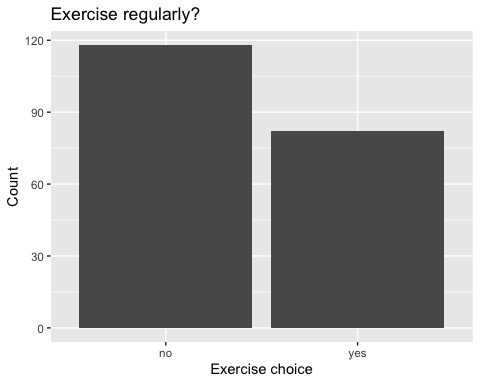


The above graph shows quite a normal distribution between the years that the participants last visited the hospital. The graph alone does not tell much but with other variable such as ethnicity or income bracket, we may be able to see who is able to go to the hospital for check up most.

# Graph 14 with variable Exercise regularly.

I will rename the column so that it doesn’t have any space in the column name.

CleanData<-  
 CleanData %>%   
 rename(Exercise = `Exercise regularly`)  
ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Exercise)) +  
 labs(title = "Exercise regularly?", x = "Exercise choice", y = "Count")

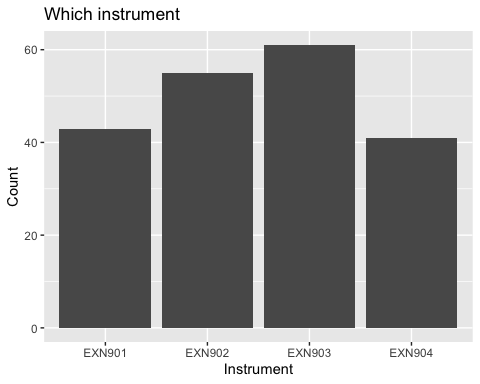


The above graph shows the 2 choices whether or not the participant exercises. It seems that the majority of people do not exercise regularly. We can also use this graph together with other variable to correlate whether or not exercising relates to health issues leading hospital visits.

# Graph 15 with variable Instrument.

This graph is probably useful for the Hematology department at the Brigham and Women’s hospital. The lab manager wants to know if samples are equally distributed to be run between 4 instruments which then give clues to why some machines breaks down more than the other.

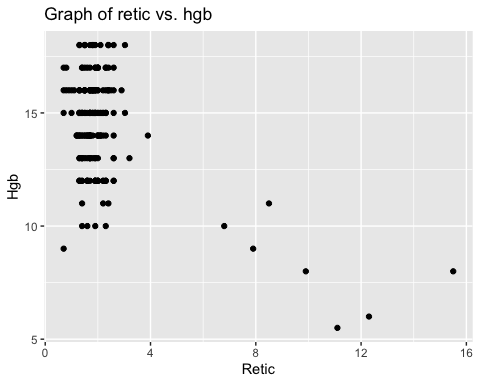
ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Instrument)) +  
 labs(title = "Which instrument", x = "Instrument", y = "Count")



The above graph shows that the 4 instruments are fairly used with similar frequency. Although instrument 2 and 3 are slightly used more frequently compared to the rest. This information can help the manager figure out the lab’s dynamics and propose solution to help distribute the use of instruments more evenly.

# Graph 1 with 2 quantitative variables: retic vs. hemoglobin

ggplot(data = CleanData) +  
 geom\_point(mapping = aes(x = Retic, y = Hemoglobin)) +  
 labs(title = "Graph of retic vs. hgb", x = "Retic", y = "Hgb")

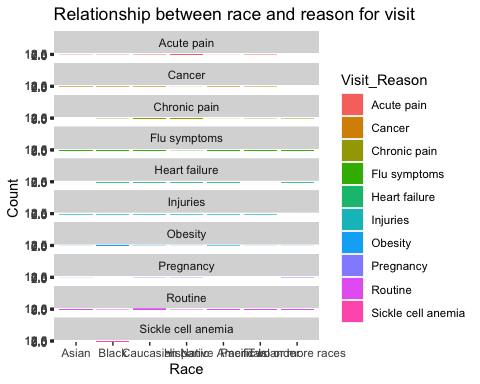


The above graph may look alittle strange and may not make much sense at first. However, if we look back at the normal range of hemoglobin and retic for adults, the main cluster of the graph represent the majority of normal values between hemoglobin and retic in adults. Hemoglobin is the protein in red blood cells responsible for carrying oxygen. Retics are immature red blood cells that are produced in the bone marrow and released in the blood stream. So, if we see a high value of retic in blood, it means that your body is actively trying to produce more red blood cells than usual. And usually this could mean that your body is trying compensate for the loss of blood or the lack of red blood cells to carry oxygen, other some other underlying issue. We can observe from the graph above that at the high values of retic, the values of hemoglobin are also lower than the normal range. This further proves that when your body is deprived of oxygen due to the lack red blood cells for example, the response mechanism would be to produce more red blood cells by pushing out early immature ones. Hence, the high values of retic.

# Graph 2 with 2 categorical variables.

Let us try to see if there is a pattern relating to race and visit reason.

ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Race, fill = Visit\_Reason)) +  
 facet\_wrap(~ Visit\_Reason, ncol=1)+  
 labs(title = "Relationship between race and reason for visit", x = "Race", y = "Count")



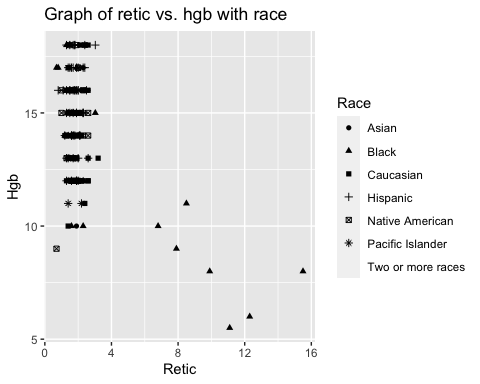
The above facet wrap chart shows the what ethnicity usually goes for hospital visit for what reason. Although the sample size of 200 is still very little to say much about the general population, we can still say one thing is that sickle cell anemia is only seen in the black population for this data set. According to the CDC statistics on 100,000 people in the United States in 2023, 90% of people are African American, and 3%-9% are Hispanic or people of Middle Eastern. So, the graph above is still a small scale compared to the large data set from the CDC. However, it still proves that the statistic is true, that is most cases of sickle cell anemia affect African American. This is thought to be due to the response mechanism to malaria, a disease with parasites transmitted by mosquitoes. However, due to the ease of travel, the trait of sickle cell anemia which is passed on genetically have migrated and spread through out continents. Statistics have been changing. Therefore, just because it is rare or never seen before, it is still better to screen for the disease when a patient is showing symptoms.

# Graph 3 of 2 quantitative variables with categorical variable as fill.

ggplot(data = CleanData) +  
 geom\_point(mapping = aes(x = Retic, y = Hemoglobin, shape = Race)) +  
 labs(title = "Graph of retic vs. hgb with race", x = "Retic", y = "Hgb")

## Warning: The shape palette can deal with a maximum of 6 discrete values because  
## more than 6 becomes difficult to discriminate; you have 7. Consider  
## specifying shapes manually if you must have them.

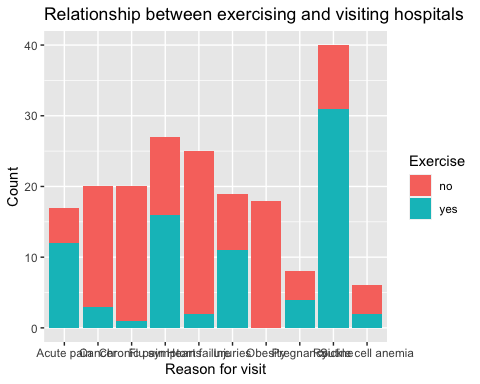
## Warning: Removed 24 rows containing missing values (`geom\_point()`).



The above graph shows the relationship between retic and hemoglobin clearer with race. Earlier we pointed out that high retic usually accompanied by low hemoglobin values. In this graph, we can see one more thing is that those outliers belong to the ethnic group “black”. We mentioned above that African American contributes 90% of sickle cell anemia cases in the U.S, and that in our dataset, 100% of participants with sickle cell anemia are black. Therefore, we can learn from this graph is that the sickle cell anemia symptoms usually include high retic and low hemoglobin values. This makes sense because the word “anemia” means not having enough healthy red blood cells or hemoglobin to carry oxygen. Hence, we observe low hemoglobin values in those patients. Their body responds with pushing out alot of immature red blood cells to compensate for the anemia symptoms in which we then observe the increase in retic values.

# Graph 4 to see the relationship between visiting hospital and exercising

ggplot(data = CleanData) +  
 geom\_bar(mapping = aes(x = Visit\_Reason, fill = Exercise)) +  
 labs(title = "Relationship between exercising and visiting hospitals", x = "Reason for visit", y = "Count")



The above graphs clearly shows the importance of exercising regularly. Just to keep in mind that this data set is small and cannot account for the entire general population. However, just based on this dataset, we can say that patients who are obese generally do not exercise regularly. The patients who come for routine check up regularly exercise. This may mean that these patients are quite health conscious in their lifestyle. We can also see that majority of people who come for cancer, chronic pain or heart failure do not exercise regularly.