Extra Question:
What do you think ***mutate_at*** *does?*
Ans. Mutate Function in R (mutate, mutate_all and mutate_at) is used to create new variable or column to the dataframe in R. Dplyr package in R is provided with mutate(), mutate_all() and mutate_at() function which creates the new variable to the dataframe.

1. **Data exploration:** To explore visually whether blood pressure levels and total cholesterol levels are associated with heart disease, create boxplots of *sysBP*, *diaBP*, and *totChol*, broken up by the levels of *TenYearCHD*. [**Hint:** Dynamic plots may help understanding!]

2. **Data preprocessing:**

(i) Read the data file into R. Set the seed to **123** and split the data into `dffTrain` and `dffTest`. Randomly sample 70% of the data for training, and use the rest as test dataset.

(ii) What are the proportions by gender in your training vs. test set? How does the distribution of age look? Looking at these, do you observe any signs of a sampling bias?

Hints:

[A] It's time to use R like a pro! You can pipe your *dffTrain* into the `group_by(variable)` function and then into **`tally()`** *-no arguments-* to get the counts across a group.

- To add percentages, pipe one more step into `mutate(pct = 100*n/sum(n))`

[B] For a continuous variable like age, there are so many groups, right? Each age is practically a different group. In such cases, you may want to create your own groups.

- You can use *`ageGroup=cut_interval(age, length=10)`* in group_by()

[C] You can also create a histogram for age, which probably makes more sense.

- After creating the histogram, try adding `fill=gender` into aes() of ggplot(), and see what happens. In addition, define `color='black'` inside the histogram!

Ans. First of all, the testing data shows that there are around 44% males and 56% females. When compare it with the training data set, it is observed that the proportions are not very variant. Secondly, for the distribution of age, it can be seen that the percentage distribution for each age group in both the testing and training data set are similar. So, there is not a lot of signs of a sampling bias when it comes to these 2 factors.

3. **Linear probability model:** Build a linear probability model `fitLPM` using all variables in `dffTrain`. Make sure to check for collinearity by both thinking about the variables, and

using VIF values as guiding signals, and take necessary precautions. You know how to mitigate collinearity (if not, please ask during the lab!). After finalizing the model, which of the variables are statistically significant at the 95% level? What does this model tell you about the risk factors of heart disease? Do you have any reservations? Discuss.

Hints:

[A] To include all the variables, use a full stop `.` To exclude a variable, use a negative `-`

[B] Run diagnostics to see whether this model violates the linear regression assumptions.

Ans. We do not know what relation is there, between diaBP and sysBP so we cannot drop them even if their VIF values are high. But, currentSmoker and cigsPerDay show a strong multicollinearity. The variable currentSmoker is dropped as it tells us if the at this point person smokes or doesn't. So, I drop it. Then, when we finalize the model, we get variables that are significant at 95% level as, gender1, age, cigsPerDay, sysBP, glucose, prevalentStroke1, prevalentHyp1, heartrate. Other variables which exist in the model and are not significant can be removed from it so as to improve the multiple R-square(variance) value of the model. The variables influencing the chronic heart disease can be figured out in the model.

4. Speaking of using R like a pro, a better way to run a model and create a results table with predictions is as follows. Please run this code to make predictions using the LPM model and store them into *resultsLPM*

```
resultsLPM <-
    lm( …fill in here… ) %>%
    predict( …fill in here… ) %>%        => Use the option type='response' for probabilities
    bind_cols(dffTest, predictedProb=.) %>%        => The dot marks where to pipe into
    mutate(predictedClass = …fill in here… )        => Use 50% as cutoff for classification
```

Inspect `resultsLPM`. Then, **copy and paste your code from Q2-ii** and check the prevalence of *TenYearCHD* in the *test dataset* this time. How many people have heart disease in reality (in the test dataset)? Run the same code for *predictedClass* in the *resultsLPM*. How many people did the model predict having heart disease? Compare and report your observations.

Before you continue:

You may have noticed that we did not convert TenYearCHD into a factor yet, even though it is a factor. This is because we wanted to use it in a linear model. It is time to make it a factor.

- Use `mutate()` to convert TenYearCHD to a factor both in *dffTrain* and *dffTest* datasets.

Ans. 172 people have heart diseases in reality, that is, in the test data set. After running the predictedClass and resultsLPM, it is seen that model predicts only 10 people having heart diseases out of 172 cases. predictedClass says that 1087 cases have no heart disease but there are 925 people in the reality that do not have a heart disease. There is some significant difference in the predictions here.

5. **Logistic regression:** Build a logistic regression using the predictor variables you decided to keep in the model you built in Q3. Which variables are statistically significant at the 95% level? Compare your results with the results you obtained from the model in Q3.

Hint: See the appendix for an annotated logistic regression output in R with the definitions.

Ans. The variables gender1, age, cigsPerDay, prevalentHyp1, totChol, sysBp, heartrate, glucose are statistically significant at 95% level since they have low p value as compared to the alpha value which is 0.05. When we compare the results with the model of Q3, we see that prevalentStroke1 is not present in new model. That is why, prevalentStroke1 is not statistically significant.

Interpret the following variables: *age*, *gender*, and *diabetes* (whether significant or not):

- **Hint:** You can run **`exp(coef(fit))`** after a logistic model to exponentiate the coefficients of all variables at once, and use them in your interpretations.
- **Type these interpretations AFTER completing the lab unless you have any questions.**

Age: **On average, when everything else is constant, as age increases by one year then odds of having a heart disease increases by 6.9 %.**

Gender: **On average, when everything else is constant, male gender will have a probability of heart disease by 1.525 times more than females.**

Diabetes: **On average, when everything else is constant, person having heart disease and diabetes will have 0.995 times more than a person not having diabetes.**

6. Create a new results table ***resultsLog*** by using the logistic model. Let's continue like a pro.

Hint: You will follow the same steps you took in Q4 but this time for logistic regression. This means, **your predictedClass will need to be defined as a factor** (you know how to do this!).

How many people did the logistic model predict having heart disease? Report your observations and compare them with the actual values, and the predictions of the linear probability model from Q4. Do you think the logistic model is an improvement? Why?

Hint: For now, continue to use your code from Q2-ii to create the tables for comparison.

Ans. According to the dataset, there are 172 people who have a heart disease. But, the model here predicts only 19 people to have a heart disease. The difference is so big in this context here i.e. 172 and 19. This is a problem and hence it is not a good prediction. Well, in this model's defence it predicts more cases than the linear model which predicted only 10 cases which tells us that it is comparatively a better model. But then, it is still nowhere enough because it should predict a lot more.

7. It is time to create a confusion matrix, a final step before evaluating performance (which we will cover next week). As you're using R like a pro, it is so easy to create a confusion matrix.

- Pipe the *resultsLog* dataframe you created in **Q6** into the function `conf_mat(truth = ..., estimate = ...)`
- **Optional:** Pipe one more step into `autoplot(type = 'heatmap')` to color code. This is useful when more than two classes are involved. For now, this is just a learning point.

Explain what the matrix tells you in addition to what you learned from the tables in **Q6**.

Ans. Confusion matrix values that we see are:

TN : 919
TP : 13
FN : 159
FP : 6

Our confusion matrix predicts if a given person in the dataset has a Coronary heart disease or not. It is important to predict the output correctly as it is about a person's life here. So, if we predict the false negative in high quantity then it is a problem because it will mean that the person does not have a Coronary heart disease despite of having one. And, here exactly that is happening. The value of FN is pretty high which is dangerous. Ideally, we should see the true positive values to be high because that means we predict correctly all those with a positive result. But here, that value is low which means there is a problem with this confusion matrix.

8. No analysis is complete without a visualization. Plot the relationship between the statistically significant variables (*age*, *cigsPerDay*, *totChol*, *glucose*) and the probability of heart disease:

- Note that you stored the predicted probabilities as *predictedProb* in the *resultsLog* in Q6.

- Use `geom_point()` and `geom_smooth()` after `ggplot()`, without adding any parameters
- Be creative. For example, add `color=currentSmoker` (or `=gender`) into the `aes()`
- Add a title for the plots, and label both axes [**Hint:** You can use the labs() function]

Discuss your observations.

Ans.

Age vs Probability of heart disease:

By observing the graph, it is evident that both men's and women's probability of having a heart disease increases with increase in age. This can be seen because of a steady increase in our graph. But, the curve for men is above the curve for women which indicates that on average, men always bear a higher chance of getting a heart disease when compared to that of women.

Cigarettes per day vs probability of heart diseases:
In males, the probability of risk of heart disease increases at a slow rate when the male smokes between 0 to 40 cigarettes a day. However, the probability starts increasing sharply after 40 cigarettes. I don't exactly know why exactly after 40, but smoking so many cigarettes (40+) is bound to increase the risk of heart diseases since it will cause heavy damage to the body. When it comes to females, there is not a particular consistent trend. We see that the probability of a heart disease decreases when a female smokes between 0 to 10 cigarettes a day. Then if the number of cigarettes is increased to 10 to 20, the probability increases as expected. The probability remains around the same mark if the number of cigarettes smoked are increased to 20 to 45. The probability of heart disease for a male is more than that for a female. Females have the graph ending at 45 whereas men's graph doesn't end there. It goes on till 60.

Total Cholesterol vs probability of heart disease:

There can be seen that the rate of increase of probability in men and women is different. But, this is with the fact that the trend is similar as an increase can be observed in both. Men's rate of increase is lower compared to that of women. The probability of women's heart disease starts at a lower point as compared to men(at around 0.05 as compared to 0.12) but as the rate of increase for women is high, the probability increase for women as the cholesterol levels go above approx. 315-320. Another thing to be noted is how the cholesterol levels of women go all the way to almost 380 whereas the levels of men end at 360 which indicates that women may have more cholesterol than men. This suggests an even elevated increase in the probability of a heart disease in women.

Glucose vs probability of heart disease:

For both males and females there is an upward increasing trend observed. However, at the beginning, we see a fall in probability of heart disease in women when the glucose levels range around 10-20. The probabilities for both the genders start at a similar point. But, females have a higher increase in the probability of getting a heart disease as their glucose levels increase as that of men. The graph also depicts that similar to cholesterol levels, women may have higher levels on maximum glucose compared to men. This suggests an even elevated increase in the probability of a heart disease in women.

Switching to a new framework "Caret" we will continue to use in this course from now on:

9. You already loaded the "caret" library at the beginning. If not, load it now. Replicate the analysis in Question 6, this time using the caret library. Use Appendix II for guidance.

- Name the results table resultsLogCaret and create it using the train function.
- Inspect resultsLogCaret carefully, compare it with resultsLog from Q6 and discuss.
- Create the confusion matrix using caret, and compare it with the one in Q7. Discuss.
- Don't worry about the rest of the output after the matrix. We will discuss it next week!

Ans. Bearing the same predictedClass values, the results of resultsLog and resultsLogCaret are one and the same. Because we use different methods but work on the same data set, this should not come as a surprise. GLM function and Caret library are the two things used in the corresponding methods. The results for both the models are strengthened more by the fact that the confusion matrix for resultLogCaret is same to the values in the question 7.

10. Now that you have learned how to use logistic regression for classification, and how to do so **using the caret library**, you can solve another business problem for *Banco Portugal*. See Appendix III for the details of the dataset. The bank runs a telemarketing campaign for a savings account. Have you ever received one of those promotions by the way? "Open a savings account today and get XXX$ bonus!" See this month's promotions by clicking here.

Banco Portugal hires you to predict whether a customer will open an account. The bank will use your model to develop promotional campaigns with higher conversion rates. Load the data, make conversions of variables as you see fit, and build logistic regression models using the caret library. Explore at least three alternative models, compare their performance, and pick a final model. Show your full work in the R Notebook. Below, discuss only your findings, your final decision, and explain how your final model helps Banco Portugal with its purpose.

Now that we have discussed the performance measures, you can decide on a performance metric (or two) beyond just accuracy to compare the models and explain your reasoning. Because the caret library already reports the values of performance measures by default, you don't need to do any coding -This part is pretty much a thinking and reflecting exercise!