1. How did you clean the data and what was wrong with it? Close to 90% of a Data Scientist's job is in cleaning data.

I followed a number of steps for cleaning the data and they are listed in more technical details in the R Markdown file under stage 2 of the solution part.   
Here is the overview of the steps I followed and the issues I identified with the data.

1. I checked the data summary statistics. Everything in the summary statistics for **days\_since\_last\_order** was greater than **days\_since\_first\_order**. That can’t be true. Since first day is the day you placed first order and last day is the day you placed your last order, your first and last order can either be on the same date or your first order date is greater than last order date.
2. Once I identified first issue, I check the top few rows of the data to verify the issue.
3. I can see **days\_since\_last\_order** and **days\_since\_first\_order** both are corrupt columns.
4. I went further to investigate the extent of the data corruption. I filtered the records where **days\_since\_last\_order** were greater than **days\_since\_first\_order** and stored it in the corrupt\_records. Then I calculated the percentage of corrupt records. This provided me with the insight that almost 94% of the data is corrupt.
5. I went further and viewed the data and figured out that noise was introduced by multiplying the **days\_since\_first\_order** with 24 and storing that value in **days\_since\_last\_order**.
6. After that I just filtered the data to get the records where number of days since first order are greater or equal to days since last order.
7. Next I checked if the orders are more than cancellations in the clean dataset. I didn’t find the issue here.
8. Next step was to use correlation plot to see if there is any correlation in different variables that can influence the result. I didn’t find any significant correlation that I should worry about.
9. What are the features you used as-is and which one did you engineer using the given ones? What do they mean in the real world?
   1. The features that I used as in are:
      1. Days\_since\_first\_order
      2. Days\_since\_last\_order
      3. Average\_discount\_onoffer
      4. Average\_discount\_used

The day features are useful to understand the customer loyalty. The discount features can tell us whether discount have any impact on user purchasing behaviour. The discount was on female items, male items or unisex items and how it impacted customer return to site ratio.

* 1. The features that I engineered are:
     1. FemaleItemsRatio
     2. MaleItemsRatio
     3. UnisexItemsRatio
     4. WappItemsRatio
     5. MaccItemsRatio
     6. WftwItemsRatio
     7. MappItemsRatio
     8. WaccItemsRatio
     9. MftwItemsRatio
     10. WsptItemsRatio
     11. MsptItemsRatio
     12. CurvyItemsRatio
     13. SaccItemsRatio
     14. MsiteOrdersRatio
     15. DesktopOrdersRatio
     16. AndroidOrdersRatio
     17. IosOrdersRatio
     18. OtherDeviceOrdersRatio
     19. WorkOrdersRatio
     20. HomeOrdersRatio
     21. ParcelPointOrdersRatio
     22. OtherCollectionOrdersRatio
     23. RevenuePerOrder
     24. ItemsPerOrder
     25. CancellationRatio
     26. ReturnRatio
     27. VoucherRatio

The items ratios are created using total number of items and specific items group numbers. These give us idea around the type of items mostly order on our website.

The device specific ratios are created using total orders and device specific order numbers. To see which group of people using what type of device to place the order.

Delivery type ratios are created using total orders and delivery specific order numbers. How different segments might be using delivery options such as other collection or parcel.

Delivery ratios are created using total orders and delivery specific order numbers for home and work. If the people are further opting for more of work delivery or home delivery. If work delivery is more dominant, we might think of having our warehouse close to CBDs and have more stock in them to cut the delivery cost.

Revenue per Order, return ratio, cancellation ratio and voucher ratios are all based on the number of orders. This give insight about the ordering behaviour.

1. What does the output look like - how close is the accuracy of the prediction in light of data with labelled flags?  
     
   I used the random forest and the accuracy is around 99 percent on the validation set. The Important features identified are FemaleItemsRatio, MaleItemsRatio, WappItemsRatio, MappItemsRatio and MftwItemsRatio.
2. What other features and variables can you think of, that can make this process more robust? Can you make a recommendation of top 5 features you'd seek to find apart from the ones given here?  
     
   From the model, we can see that FemaleItemsRatio, MaleItemsRatio, WappItemsRatio, MappItemsRatio and MftwItemsRatio are the key features in determining user gender. So, I will need these features to be part of any future dataset. The purchasing behaviour and amount spent is not really helping in determining the gender.  
     
   I will look for more randomized data first. Around 90 percent of the data was part of one major segment. This will bias the model and create false positives if the data is spread more evenly let say 60-40.