Ten questions an audience would ask with answers:

1) Where did you get this data and what are the limitations?
    - This data is from Kaggle.com and is titled Health Insurance Premium charges based on Gender, BMI and other characteristics. The limitations of the data are that it is not verified in terms of its source, so it is not known if the data is real or a sample of an insurance database, etc. Because of this limitation, any result will need to be replicated prior to making viable interpretations.
2) What other analyses could you do with this data or is what you did it?
    - In terms of other regression analysis, I covered linear regression, PCA and random forest regression. Another analysis method that I could have conducted would have been gradient boosting. This analysis is very similar mathematically to random forest, in that it is an ensemble method that creates many decision trees. The difference between the two is that gradient boosting builds decision trees one at a time, rather than independently, which corrects errors made by previous trees.
3) Would you consider combining this data with other insurance information to test your results?
    - This is a great comment, and it is something that I would be very interested in doing. I like the idea of taking slightly disparate data and finding the common thread on how it can be used together to tell a wider story or show an unexpected angle. It would be possible to combine this data with google trends data for health care related questions and see if you could find some relationship between the number of searches for medical questions and the amount of costs. There are many different avenues where it would be possible to combine this data with government data or more detailed insurance information to produce broader and more generalizable results.
4) Does this data have balanced demographic and health information?

- This data is balanced. Columns like BMI have bell shaped distributions while variables like Age show there are more younger people in this dataset. Conversely, those who smoke may seem imbalanced, but the percentage of smokers is higher (25%) than the representation in the general population (11%). Both Region and Gender were balanced.

5) There are many warnings against smoking, how do you think a model like this might help with getting people to stop smoking?
    - The model results approach another angle to smoking, which is the resulting medical costs associated with smoking. Additional variables that would be interesting to include would be length of time of smoking to be able to interpret when costs start to accelerate for smokers. Intervention messages could be targeted at young people about avoiding costly medical bills. To be honest, though there have been many ad campaigns against smoking, it is the rare campaign that simply shows the worsening of health indicators due to smoking. I do think this type of information can be influential and impactful as part of a health intervention.
6) You mentioned interventions but did not discuss it further, what do you think might work for this?
    - Expanding on what I previously answered, I think it would be important to list the top health costs and the annual costs spent by smokers and non-smokers. Likewise, I think it would be great to better understand potential impacts of smoking on things like exercise and diet and determine if additional interventions could be developed around these to deter smoking.
7) Did you consider using government data instead with this model? Would you expand your work in this arena to include that?
    - Government data would be essential to enhance this model – often times, government data (like large surveys collected across the United States) is released by year and region, it would be ideal to combine with data or medical claims data to show a larger picture of health and health costs in the United States.

8) Why did you choose to keep what seems to be outliers in your data?
    - The box plots in my presentation for Medical Costs and BMI had some outlier values in the data. It is important to ensure that the input data to our models is indicative to real life situations. Some individuals have extreme medical costs for whatever reasons and some individuals have higher weights. It is important to include actual values even if those deviate from the mean (average) in a dataset where they are part of the population and not necessarily highly unusual.
9) Were there other relevant relationships with this data that you felt you did not explore?
    - This data presented a limited number of health characteristics and demographics of subjects. It would be interesting to do additional EDA on this data and further explore the relationship between each variable with another in the data. Though I did a significant amount of EDA, I did not do every combination of variables or look at all continuous variables by smoking status (as an example). These could be additional analysis or likely a better route would be to get more robust data and see if the findings here could be replicated.
10) Do you think it would be possible to combine this data with hospital discharge data to tell a larger story of health care costs?
    - This would be ideal and something that would be very interesting and useful to pursue. This way, we could have more detailed information on the healthcare costs and setting of those services and see if there are more relationships between these variables that this modeling effort was not able to uncover.