Chak Wong
Cota Data Analyst Interview

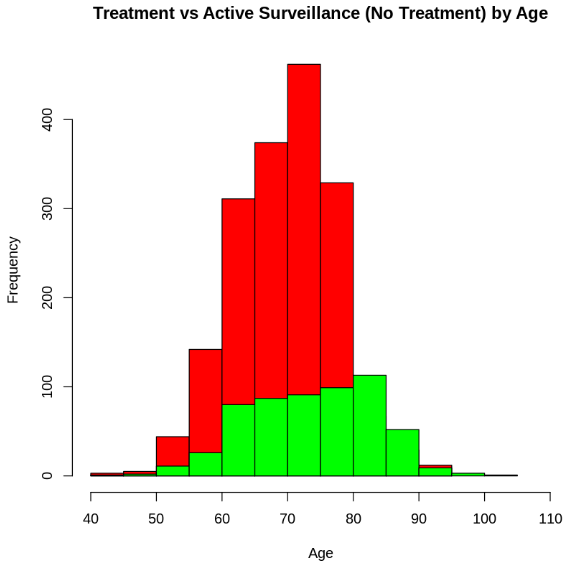
1. Perusing through the simulated_patients.csv dataset, I eliminated observations where data of final observation was beyond the scope of the data collection (1/1/07 - 1/1/12) and filtered out duplicate entries.

2. Crosstabs illustrates the patient's gender distribution across ethnicity - generally, there appears to be a ~2:1 ratio of female: male. However, chi-square test of independence finds that the gender distribution isn't associated with ethnicity. I eyeballed the gender ratio across ethnicity. To calculate the chi-square, I converted the two-way frequency table before running chi-square test.

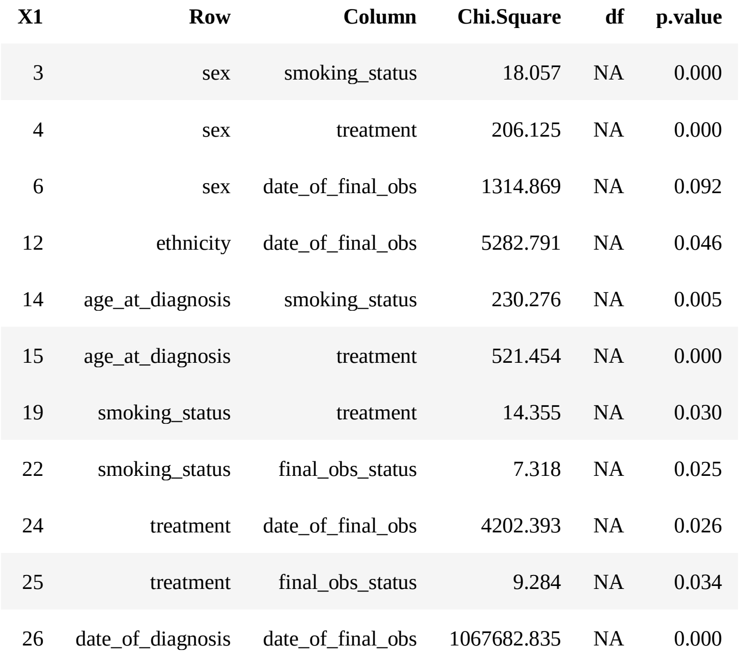
```
        ethnicity
sex      Asian or Pacific Islander Black (not Hispanic) Hispanic
  female                       246                  147      243
  male                         179                   75      156
        ethnicity
sex      Native American White (not Hispanic)
  female              40                  718
  male                29                  536

          Pearson's Chi-squared test

data:  sex_by_eth_tbl
X-squared = 7.1685, df = 4, p-value = 0.1272
```

3. Crosstabs illustrates the patient's age [of diagnosis] distribution by treatment. Initial Chi-square test of independence finds that the age distribution is associated with being on active surveillance, though sample size across groups were too small. A follow-up chi-square test of independence using simulated p-values, though, suggests age distribution is associated with being on active surveillance. Specifically, older patients [at the time of diagnosis] have a higher chance of being on active surveillance (no treatment). To calculate this, I first renamed the factor levels in treatment to address the question before observing the two-way frequency table, and then running the chi-square test, similar to the strategy in question 2.

```
Pearson's Chi-squared test with simulated p-value (based on 2000
          replicates)

data:  age_by_treat_tbl
X-squared = 344.39, df = NA, p-value = 0.0004998
```



4. Running chi-square tests between all possible combinations of patient characteristics, there are very strong relationships (p < .005) between sex and smoking status, sex and treatment, age_at_diagnosis and smoking status, age_at_diagnosis and treatment, and date of diagnosis and date of final observation, and strong relationships (p <.05) between ethnicity and date of final observation, smokine status and treatment, smoking status and final observation status, and treatment and date of final observation. It should be noted that simulations were run due to small sample sizes. Fisher's exact test didn't work with the function. To address this question, first removed the patient_id and patient name from the dataframe before using a function to obtain all possible combinations for a chi-square test. For complete table, refer to code attachment.



X1	Row	Column	Chi.Square	df	p.value
3	sex	smoking_status	18.057	NA	0.000
4	sex	treatment	206.125	NA	0.000
6	sex	date_of_final_obs	1314.869	NA	0.092
12	ethnicity	date_of_final_obs	5282.791	NA	0.046
14	age_at_diagnosis	smoking_status	230.276	NA	0.005
15	age_at_diagnosis	treatment	521.454	NA	0.000
19	smoking_status	treatment	14.355	NA	0.030
22	smoking_status	final_obs_status	7.318	NA	0.025
24	treatment	date_of_final_obs	4202.393	NA	0.026
25	treatment	final_obs_status	9.284	NA	0.034
26	date_of_diagnosis	date_of_final_obs	1067682.835	NA	0.000

5. The average amount charged for patients who are 65 years or older at diagnosis is $88,371.20. Among those charged $100,100 or more, about 83% are patients who are 65 years or older at diagnosis. To address this question, I first used a function to keep only the numeric portion of the alphanumeric string for each patient, before merging the two datasets together by patient_id; for the revenue dataset, I obtained the sum charge by patient_id. Second, I filtered the newly merged set and obtained the average sum charge per patient who are 65 years old or older. Third, to obtain the proportion who are age 65 or older among those charged at least $100,000, I first filtered the dataset to keep those with at least $100,000, before obtain the ratio of those 65 years old or older over the everyone who was charged $100, 000.

6. Minus patient id and patient name, and date of and status at final observation (assuming physician's choice of therapy was at the date of diagnosis), the model was run on the sub-setted dataset to discern what factors, if any, were associated with the physician's choice of targeted therapy vs chemotherapy. Results indicates sex, age at diagnosis, and date of diagnosis as having significant effect on the doctor's choice of therapy, while ethnicity, smoking status, and total charge amount were found to be insignificant. Results, however, should be considered with a caveat given the skewed and kurtotic nature of the distribution in the QQ-plot. To address this question, I first filtered treatment variable to keep only chemotherapy and targeted therapy, before running the reduced model in a logistic regression, keeping only variables chronologically before a doctor’s diagnosis.

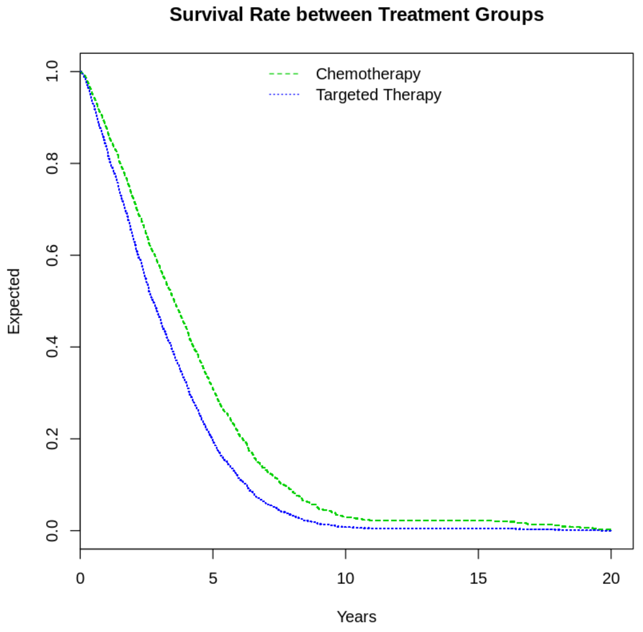
```
Call:
glm(formula = treatment ~ sex + ethnicity + age_at_diagnosis + 
    smoking_status + date_of_diagnosis + sum_charge_amount, family = "binomial", 
    data = Q6_filtered_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0760  -0.9131  -0.5333   1.0064   2.2435  

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -1.071e+01  1.818e+00  -5.889  3.9e-09 ***
sexmale                       -1.525e+00  1.134e-01 -13.447  < 2e-16 ***
ethnicityBlack (not Hispanic) -2.202e-01  2.171e-01  -1.014    0.310    
ethnicityHispanic             -1.354e-01  1.813e-01  -0.747    0.455    
ethnicityNative American      -4.783e-01  3.653e-01  -1.309    0.190    
ethnicityWhite (not Hispanic) -9.531e-02  1.460e-01  -0.653    0.514    
age_at_diagnosis              -8.173e-02  7.686e-03 -10.634  < 2e-16 ***
smoking_statusnon-smoker       7.497e-01  5.609e-01   1.336    0.181    
smoking_statussmoker          -4.831e-01  6.066e-01  -0.796    0.426    
date_of_diagnosis              1.278e-08  1.389e-09   9.199  < 2e-16 ***
sum_charge_amount             -1.456e-06  1.746e-06  -0.834    0.404    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2425.5  on 1777  degrees of freedom
Residual deviance: 2061.2  on 1767  degrees of freedom
AIC: 2083.2

Number of Fisher Scoring iterations: 4
```

7. Among the sample who expired, the median time to death is 1072.5 days (or ~2.94 years). To address this question, I filtered the dataset to include only patients who expired (i.e. died), before calculating the time difference between date of last observation and date of diagnosis, and calculating the median of this vector.

8. On average, there is a 48% probability that a patient will be alive after 1 year. To address this, I ran a survival curve, and calculated the mean p-values of all the patients.

9. Yes, survival is different for patients treated with chemotherapy vs. patients treated with targeted therapy. Generally, patients under chemotherapy are ~1.39 times more likely to live longer than patients treated with target therapy, though the diffrence in probability between the two treatment groups varies across time. For example, according to the plot, at five years the expected survival rate for individuals under chemotherapy is ~30% while those under targeted therapy is ~30%. To address this question, I filtered out the dataset to include only chemotherapy and targeted therapy, dropping out the empty factor levels, before running cox regression on the survival curve using only variables found to be significant from the logistic regression problem (i.e. treatment, sex, age_at_diagnosis, and date_of_diagnosis). To see whether the two treatment groups were significantly different from each other, I ran an anova on the `coxph(surv))` output.

```
Call:
coxph(formula = Surv(time_to_death, final_obs_status_num) ~ treatment, 
    data = Q9_data)

  n= 1778, number of events= 1679 

                            coef exp(coef) se(coef)     z Pr(>|z|)    
treatmentTargeted Therapy 0.32863   1.38906  0.04955 6.632 3.31e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                          exp(coef) exp(-coef) lower .95 upper .95
treatmentTargeted Therapy     1.389     0.7199      1.26     1.531

Concordance= 0.547  (se = 0.007 )
Rsquare= 0.024   (max possible= 1 )
Likelihood ratio test= 43.19  on 1 df,   p=4.959e-11
Wald test            = 43.98  on 1 df,   p=3.313e-11
Score (logrank) test = 44.37  on 1 df,   p=2.72e-11
```


Survival Rate between Treatment Groups
Chemotherapy
Targeted Therapy
Expected
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Years