Q1

(a) Since we have that, for this problem, a sample size of 30 is considered large, the CLT would tell us the both data samples would be roughly and greatly approximated by a normal.

(b) Let s_1 and s_2 be the standard deviation for the first and second samples respectively

$$s_1 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{165}(y_{i1} - \bar{y}_1)^2} = \sqrt{\frac{1}{164}34.26} \approx 0.457$$

$$s_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{35}(y_{i2} - \bar{y}_2)^2} = \sqrt{\frac{1}{34}37.68} \approx 1.053$$

The standard deviation of the first sample is significantly less than the standard deviation of the second sample; this is due to the larger sample size of the first, leading to less variance. It would not make sense that the population standard deviation are equal, since the sample standard deviation differs greatly.

(c) The discrepancy measure / test statistic (given that we cannot assume the population standard deviations are the same) is

$$D = \frac{\left|(\bar{Y}_1 - \bar{Y}_2) - 0\right|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \implies d = \frac{|(\bar{y}_1 - \bar{y}_2)|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{|3.321 - 2.712|}{\sqrt{\frac{0.209}{165} + \frac{1.108}{35}}} \approx 3.355$$

(d) Using the conservative measure of the degrees of freedom, we have that $df = \min(164, 34) = 34$. Looking at a t table, with 34 degrees of freedom, we have that for $t \approx 3.355$ the $p-$value would be in the interval $[0, 0.001]$. Using R code we find the $p-$value to be ≈ 0.0009, which falls in our interval. This tells us that we reject the null-hypothesis, meaning it is plausible to conclude that there is a difference in the lung capacity of vapers and non-vapers.

(e) Using the formula, we have that

$$df \approx \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} = 36.76$$

Let us round up for good measure. Using $df = 37$, looking at the t table, our $p-$values lies in the interval $[0, 0.001]$. Once again, our R code verifies this and gives us a $p-$value of ≈ 0.0009

Q2

(a) The standard deviation of the log of the claims are

$$s_1 = \sqrt{\frac{1}{n-1}\sum_i (\log(y_{i1}) - \log(y_1))^2} = \sqrt{\frac{1}{11}(52.864)} \approx 2.192$$
$$s_2 = \sqrt{\frac{1}{n-1}\sum_i (\log(y_{i2}) - \log(y_2))^2} = \sqrt{\frac{1}{14}(74.526)} \approx 2.307$$

It does make sense to assume the population standard deviation of the log of the claims are equal since both sample population standard deviations are very close in value.

(b) The pooled sample standard deviation of the log of the claims can be computed via the formula

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(11)(2.192)^2 + (14)(2.307)^2}{25}} \approx 2.257$$

(c) The discrepancy measure / text statistic is

$$D = \frac{\left|(\overline{\log(Y_1)} - \overline{\log(Y_2)}) - 0\right|}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \implies d = \frac{\left|\overline{\log(y_1)} - \overline{\log(y_2)}\right|}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{|5.967 - 5.665|}{2.257\sqrt{\frac{1}{12}+\frac{1}{15}}} \approx 0.345$$

(d) For a pooled sample population, the degrees of freedom are $df = n_1 + n_2 - 2 = 25$. Using a t table, we find that our $p-$value is greater than 0.10. Using R, we find a $p-$value of ≈ 0.366. Thus, we do not reject the null-hypothesis.

Q3

(a) Using the formula, we have that

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-90.01}{\sqrt{(667.40)(28.16)}} \approx -0.657$$

(b) A linear model between x and y is not appropriate. The scatter plot (and correlation coefficient) tells us that there is a potential negative and non-linear correlation between x and y, but the standardized residuals show non-constant variance. The QQ plots of the residuals does not fit the normal distribution. It deviates heavily near the tails and has some outliers.

(c) Once again using the formula, the correlation between x and z is

$$r_{xz} = \frac{S_{xz}}{\sqrt{S_{xx}S_{zz}}} = \frac{-70.55}{\sqrt{(667.40)(14.27)}} \approx -0.723$$

(d) The scatter plot, Standardized residuals, and QQ plot, show some linear correlation; the scatter plot is less "non-linear", the variance looks slightly more, but not entirely, evenly spread about 0, and the QQ plot of the residuals follows a more normal distribution. This suggests there is slightly stronger linear model between x and $z \equiv \log y$ compared to the variables in part (a).

(e) Once again,

$$r_{xu} = \frac{S_{xu}}{\sqrt{S_{xx}S_{uu}}} = \frac{-63.31}{\sqrt{(667.40)(10.42)}} \approx -0.75$$

(f) A linear model is very appropriate given the plots. The scatter plot is perfectly linear, the variance is constant around 0, and the QQ plot fits the model well. This suggests a strong linear model between x and $u \equiv -1/y$

(g) -

Q4

(a) The coefficients are $\alpha = 4.923$ with a Std. Error of 0.165, and $\beta = 3.059$, with a Std. Error of 0.05595. The $p-$value against the null hypotheses $H_0 \ : \ \alpha = 0$ and $H_0 \ : \ \beta = 0$ is $p < 0.001$

(b) Fitting a quadratic model, the coefficients are $\alpha = 4.798$, with a Std. Error of 0.268 and a $p-$value < 0.001. $\beta = 3.085$, a Std. Error of 0.252, and a $p-$value < 0.001. γ (the coefficient of x_i^2) is -0.01175, with a Std. Error of 0.049, and a $p-$value of 0.810.

(c) The $p-$value for both α and β are small since the data closely approximates a linear relationship between x and y, and γ is large because the correlation to the quadratic term is not very large (i.e we do not reject – there is not enough evidence against – the null hypothesis that $\gamma = 0$).

(d) With the new distribution $Y_i \sim G(\mu = 5 + 3x_i + 2x_i^2, \sigma = 0.4)$, the coefficients are: $\alpha = 4.917$, with a Std. Error of 0.136 and a $p-$value of < 0.001. $\beta = 2.912$, with a Std. Error of 0.055 and a $p-$value of < 0.001. Finally, $\gamma = 2.012$, with a Std. Error of 0.045 and a $p-$value of < 0.001. These $p-$values are against the hypotheses that $H_0 \ : \ \alpha = 0$, $H_0 \ : \ \beta = 0$, and $H_0 \ : \ \gamma = 0$.

(e) We expect the $p-$value to be small because the data fits a quadratic model, therefore the coefficient must be non-zero to add the quadratic term as a parameter. So we expect to reject the null hypotheses and we states with confidence that the coefficients are all non-zero and the data follows a quadratic model.