Problem 2

Hindy Rossignol, Riya Parikh, Mrugank Pednekar, Ioannis Panagiotopoulos

2025-09-11

R setup

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(fitdistrplus)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: survival
```

Problem 2

Part a

```
# loading df
df_insurance <- read_csv("/Users/riyaparikh_computeracct/Downloads/MIT/15.072_AdvancedAn
alyticsEdge/deliverable1-analyticsedge-mit/Data/insurance.csv", show_col_types = FALSE)
```

Part b

```
mean_charges = mean(df_insurance$charges)
sd_charges = sd(df_insurance$charges)
```

Mean = 13270.42, SD = 12110.01

Part c

```
probability_between_values = pnorm(14000, mean_charges, sd_charges) − pnorm(8000, mean_c
harges, sd_charges)
```

Probability of being btwn 8k 14k based on normal distribution is 0.19.
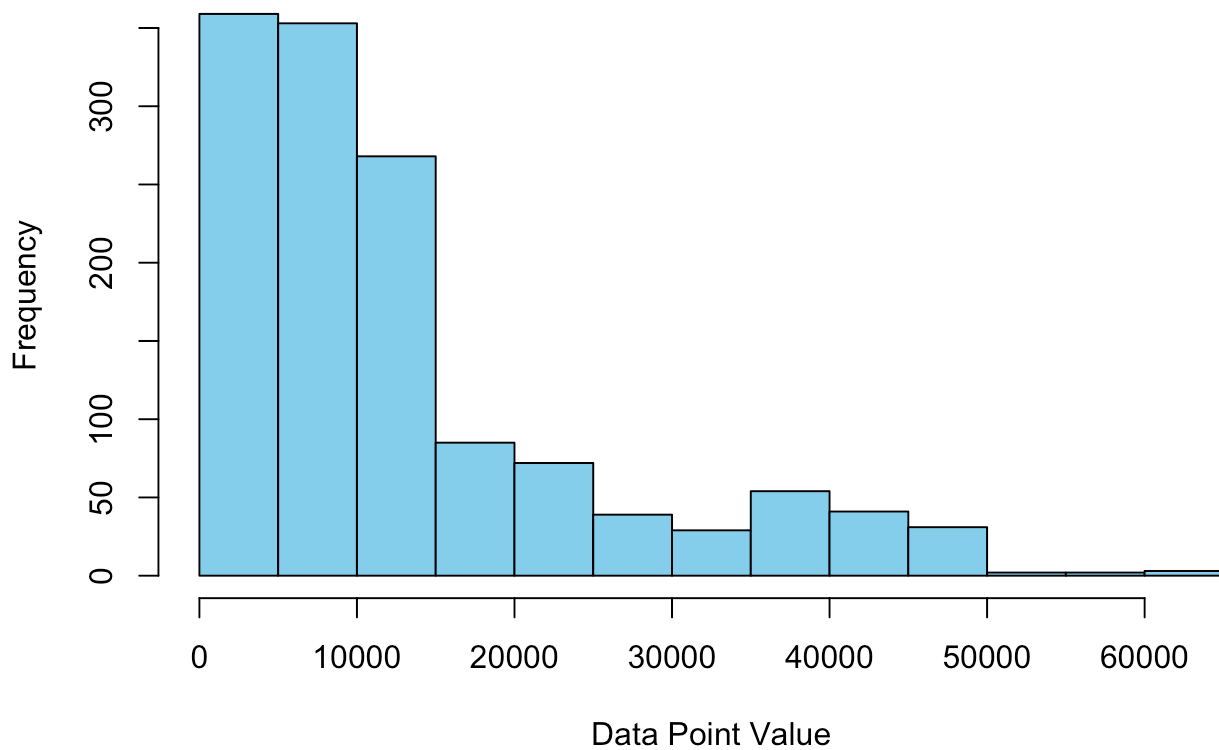
Part d

```
count_in_range <- sum(df_insurance$charges >= 8000 & df_insurance$charges <= 14000)
total_vals = length(df_insurance$charges)
true_prop_between_values = count_in_range/total_vals
median(df_insurance$charges)
```

```
## [1] 9382.033
```

```
hist(df_insurance$charges, main = "Histogram of Charges Data Points", xlab = "Data Point
Value", col = "skyblue")
```

Histogram of Charges Data Points


```
# to see overlapping stats plots - ours vs normal model
# calculate values for comparison
count_in_range <- sum(df_insurance$charges >= 8000 & df_insurance$charges <= 14000)
total_vals <- length(df_insurance$charges)
true_prop_between_values <- count_in_range / total_vals
median(df_insurance$charges)
```

```
## [1] 9382.033
```

```
# histogram of charges (empirical distribution)
hist(df_insurance$charges,
     breaks = 50,                 # more detail
     freq = FALSE,                # scale histogram to density
     main = "Skewed Charges vs Normal Model",
     xlab = "Charges",
     col = "skyblue")

# iverlay fitted normal distribution curve
x_vals <- seq(min(df_insurance$charges), max(df_insurance$charges), length.out = 1000)
y_norm <- dnorm(x_vals, mean = mean_charges, sd = sd_charges)

lines(x_vals, y_norm, col = "red", lwd = 2)

# add legend
legend("topright", legend = c("Data (Histogram)", "Normal Distribution"),
       col = c("skyblue", "red"), lwd = 2, bty = "n")
```


Skewed Charges vs Normal Model
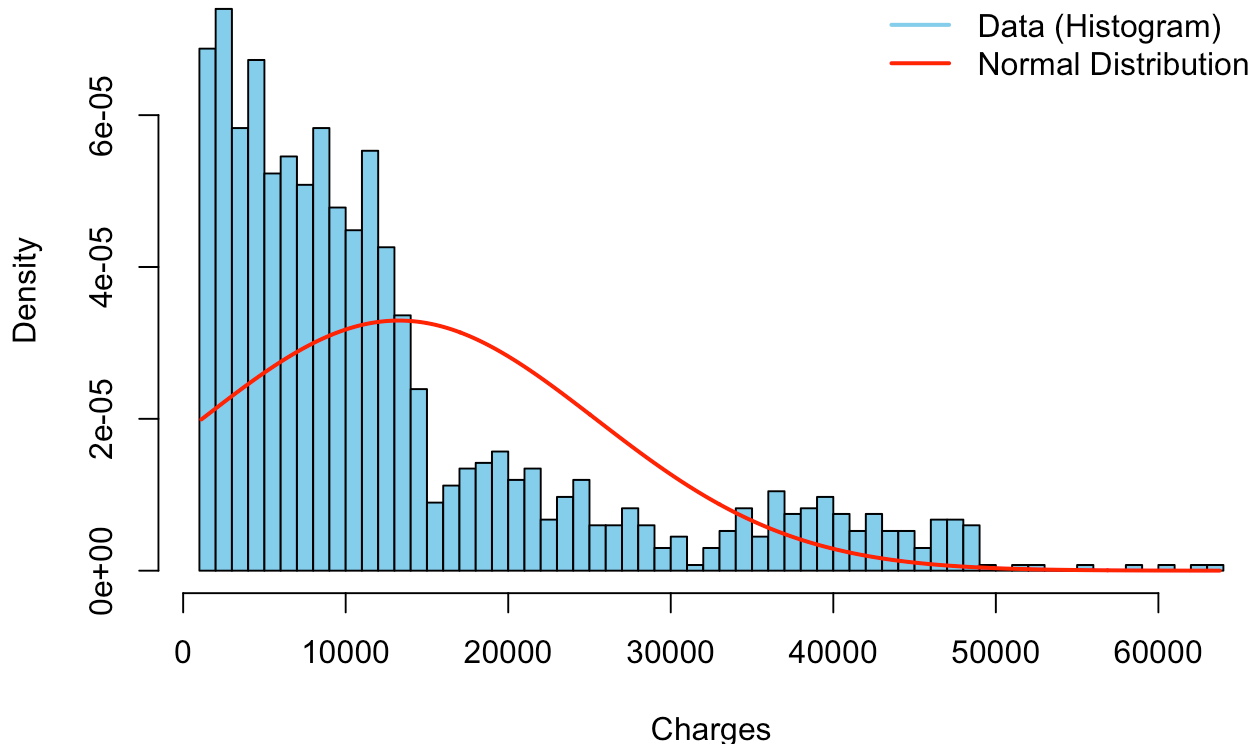
Actual proportion of charges between 8k, 14k is 0.28. This is not the same as the probability we got when assuming the charges follow a normal distribution in part c. Our initial assumption is that the charges do not follow a Normal distribution. The histogram we plotted of charges data points shows that the distribution is right skewed towards higher charges. We also see that the median charge is lower than the mean charge ($9382.033 vs $13270.42).

Part e

```
fit_norm <- fitdist(df_insurance$charges, "norm")
fit_lognorm <- fitdist(df_insurance$charges, "lnorm")
fit_unif <- fitdist(df_insurance$charges, "unif")
fit_cauchy <- fitdist(df_insurance$charges, "cauchy")

gofstat(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy))
```

```
## Goodness-of-fit statistics
##                                1-mle-norm 2-mle-lnorm  3-mle-unif 4-mle-cauchy
## Kolmogorov-Smirnov statistic    0.188462   0.0365844   0.5147556     0.185312
## Cramer-von Mises statistic     14.829729   0.3973136 155.5524079     9.501239
## Anderson-Darling statistic     85.138872   3.9424972         Inf    65.856208
##
## Goodness-of-fit criteria
##                                1-mle-norm 2-mle-lnorm 3-mle-unif 4-mle-cauchy
## Akaike's Information Criterion    28959.26    27923.58   29561.21     28716.76
## Bayesian Information Criterion    28969.66    27933.98   29571.61     28727.16
```

```
plot.legend <- c("Normal", "Lognormal", "Uniform", "Cauchy")

denscomp(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy), legendtext = plot.legend)
```
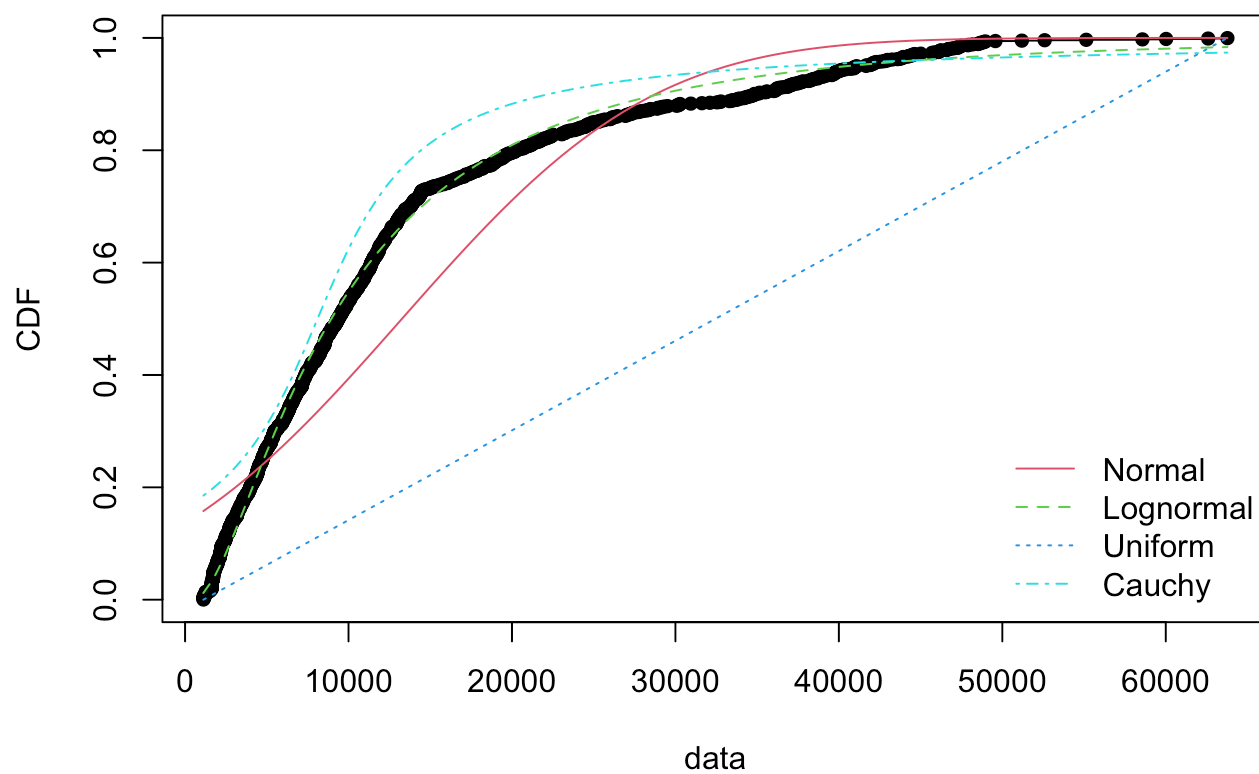
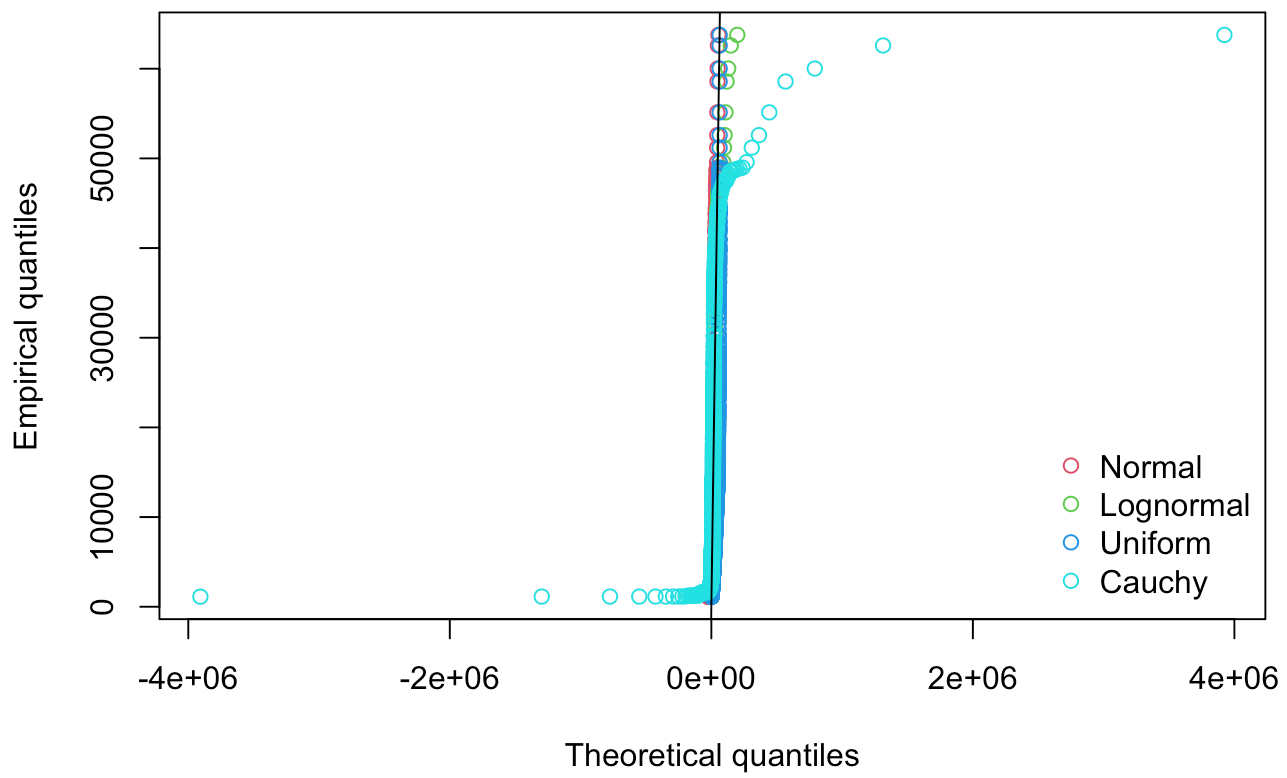

Histogram and theoretical densities
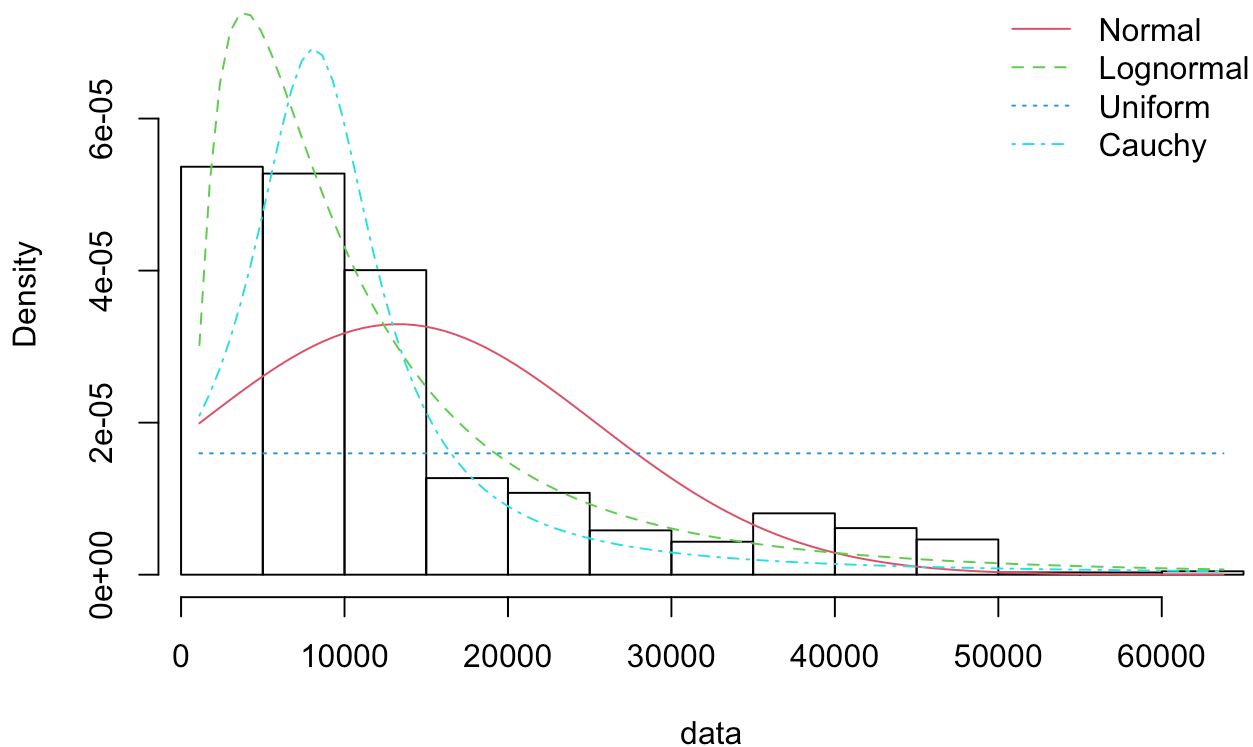
```
cdfcomp(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy), legendtext = plot.legend)
```
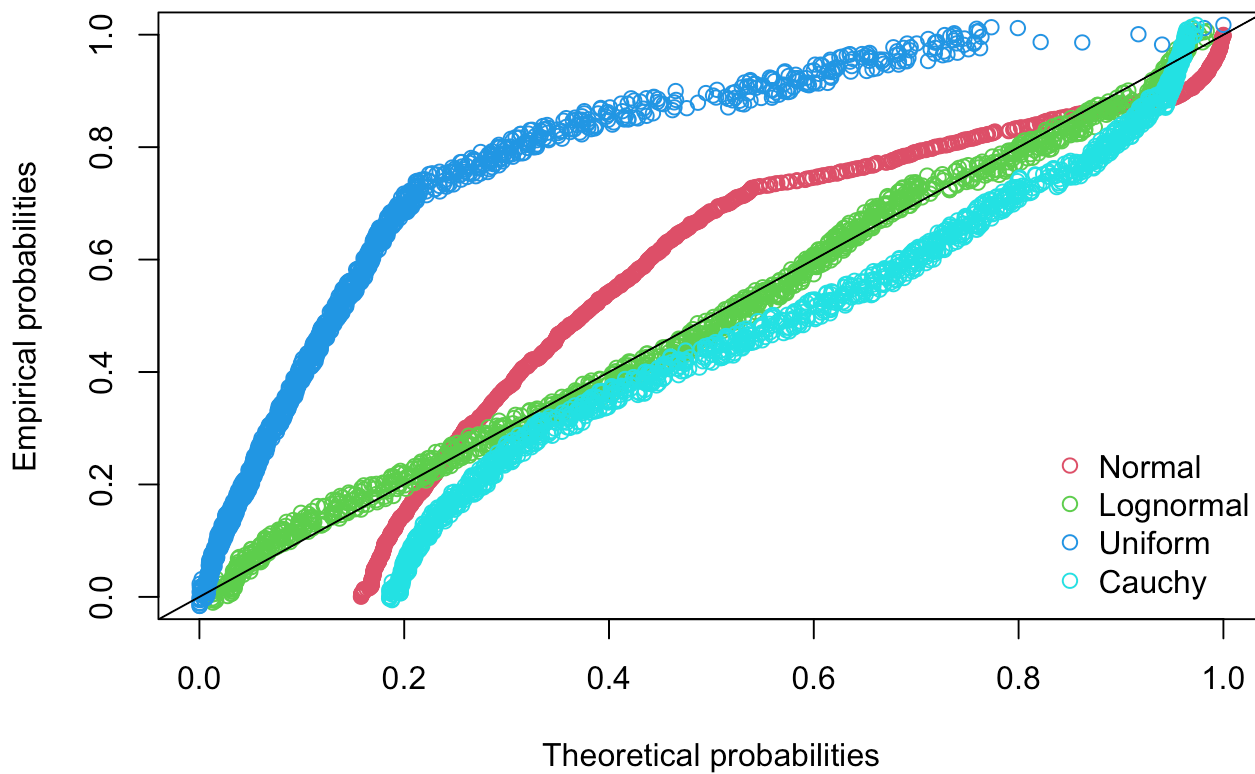
Empirical and theoretical CDFs


```
qqcomp(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy), legendtext = plot.legend)
```

Q-Q plot


```
ppcomp(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy), legendtext = plot.legend)
```

P-P plot


```
quantile(df_insurance$charges, probs = c(0.25, 0.5, 0.75))
```

```
##       25%       50%       75%
##  4740.287  9382.033 16639.913
```

```
qlnorm(c(0.25, 0.5, 0.75),
       meanlog = fit_lognorm$estimate["meanlog"],
       sdlog = fit_lognorm$estimate["sdlog"])
```

```
## [1]  4811.090  8943.289 16624.595
```

Based on both statistical metrics and visual diagnostics, the lognormal distribution is the best fit for the charges data. Lognormal wins across all metrics, especially the goodness-of-fit statistics. The KS statistic is significantly lower for lognormal, indicating a closer match between empirical and theoretical distributions. The quantile comparison shows that lognormal quantiles closely match the actual data. Refining part d, the insurance charges data is not normally distributed. Instead, it follows a log-normal distribution, which better captures the right-skew in the data and aligns more closely with observed quantiles and proportions.

Part f

```
mod <-  lm(charges ~ age + factor(sex) + bmi + children + factor(smoker) + factor(regio
n) , data = df_insurance)
summary_mod = summary(mod)
```

Part g

It is important to note that for the following interpretations, the change per unit that is being described only holds true if all other factors are held constant.

Age: for every increase of 1 yr in age, the charges are predicted to increase by $256.9, holding other variables constant. this variable is statistically significant as the pvalue is <0.05 at 2e-16.

Factor(sex)male: being a male, the charges are predicted to be lower by $131.3 compared to their equivalent female counterparts, holding other variables constant. this variable is not statistically significant as the pvalue is >0.05 at 0.693348.

BMI: for every increase of 1 unit in bmi, the charges are predicted to increase by $339.2, holding other variables constant. this variable is statistically significant as the pvalue is <0.05 at 2e-16.

Children: for every additional child the customer has, the charges are predicted to increase by $475.5, holding other variables constant. this variable is statistically significant as the pvalue is <0.05 at 0.000577.

Factor(smoker)yes: being a smoker, the charges are predicted to be higher by $23848.5 compared to their nonsmoking counterparts, holding other variables constant. this variable is statistically significant as the pvalue is <0.05 at 2e-16.

Factor(region)northwest: if a customer lives in the northwest, their charges are predicted to be lower by $353.0 holding all other variables constant. this variable is not statistically significant as the pvalue is <0.05 at 0.458769.

Factor(region)southeast: if a customer lives in the southeast, their charges are predicted to be lower by $1035.0 holding all other variables constant. this variable is statistically significant as the pvalue is <0.05 at 0.030782.

Factor(region)southwest: if a customer lives in the southwest, their charges are predicted to be lower by $960.0 holding all other variables constant. this variable is statistically significant as the pvalue is <0.05 at 0.044765.

Part h

```
mod2 <-  lm(charges ~ factor(sex) + bmi + children + factor(smoker) + factor(region) , d
ata = df_insurance)
summary_mod2 = summary(mod2)
```

The first model had an r^2 of .751 meaning that 75.1% of the variation in charges is accounted for by the model built on all the factors. the adj r^2 is .749. The second model (built without age) has an r^2 of .664 meaning that 66.4% of the variation in charges is accounted for by the model built on all the factors without age. the adj r^2 is .662. When age is not considered as an independent variable, the model is forced to "spread" the influence of age across other variables like BMI, smoking status, and region. Younger individuals are likely to be overcharged: without age, the model can't distinguish between a healthy 25-year-old and a 55-year-old with similar BMI and smoking status. Older individuals are likely to be undercharged: older individuals typically incur higher medical costs, which the model without age can't fully capture.

Part i

I would recommend applying a log transformation to the charges variable, and then redoing the analysis using the transformed variable to assess linear relationships and model assumptions more effectively.