Take Home Exam - STA2201H Applied Statistics II

Luis Correia - Student No. 1006508566

April 07th 2020

1 - Spending Behaviour

The Canadian company KOHO provides financial services, such as spending accounts and a Visa card, through an online-only app. They have come to you with a modeling problem. They are interested in understanding how spending behavior has changed in light of the recent social distancing measures put in place in Toronto. In particular, among other things, KOHO are interested in the proportion of purchases made online, how this has changed since social distancing, and how changes differ by population subgroup (age and income level).

KOHO have given you data on every transaction made by their customers in Toronto from March 1 to April 6 2020. In addition to knowing the date and monetary amount of each transaction, you also know the broad category of type of purchase, the location of the purchase, and whether it was made online or not (1 if online and 0 otherwise). You also know the age group and income level of each customer. Assume there are four age groups (18-29, 30-44, 45-64, 65+) and four income groups (<$30,000; $30,000-$59,999; $60,000-$99,999; $100,000+).

Let's assume social distancing started on March 16, when UofT went online.

(a) Introduce notation and specify a fully Bayesian model that could be used to understand the change in the proportion of online purchases since social distancing and differences by age group and income. Note I don't think there is only one right model set-up here, there are many interesting alternatives (both in terms of structure of covariates and what form of the outcome is modeled). That said, given the nature of the data, I would encourage you to formulate some sort of hierarchical model. You will need to specify and define notation for the outcome of interest and covariates, all with appropriate indexing, and specify the likelihood, group-level models, and priors. Explain how you would assess whether the average change in the proportion of online purchases is higher for 18-29 year old in the second income bracket ($30,000-$59,999) versus 45-64 year old in the same income bracket.

{*Answer.*}

Information Input

KOHO - Financial Services company is interested in understand the purchase behavior from their customers since the social distancing has started.

Variables of interest

- Y: Proportion of purchases online

Questions

1. Does this proportion has changed since the social distancing?
2. How these changes reflected within sub-groups *age* and *income level*?

Available Data

- Transaction with credit cards from March, 1st to April 6th

Data set 1

- Date of Transaction
- Amount of transaction
- Type of purchase
- Location of purchase
- Online Transaction (1=Yes, 0=No)
- Customer ID

Data set 2

This is not relevant to the problem but merely illustrative - Customer info is probably stored into a customer database, related with other data through the `customer-ID` so we will consider them in a separate and assume we have access to both data.

- Customer ID
- Age Group (4 levels)
  - 18-29
  - 30-44
  - 45-64
  - 65+
- Income Group Level (4 levels)
  - less than $30,000
  - $30,000-$59,999
  - $60,000-$99,999
  - $100,000+

The Model

For the current study, if we consider the event *customer has done an online transaction*, what we need to model is the **proportion of success** of a Bernoulli variable. In this situation I will propose for this type of problem a **Logistic Model** with the following structure:

$$y_i|\pi_i \sim Bern\left(\pi_i\right), \text{ with } i = 1, \ldots, n$$

$$\pi_i = logit^{-1}\left(\beta_o + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \alpha_{m[i]}^{income} + \alpha_{g[i]}^{age}\right)$$

$$\alpha_m^{incomel} \sim N\Big(\mu_{inc}, \sigma_{inc}^2\Big), \text{ for } m = 1, \ldots, 4$$

$$\alpha_g^{age} \sim N\Big(\mu_{age}, \sigma_{age}^2\Big), \text{ for } g = 1, \ldots, 4$$

Suggested priors

$$\mu_{inc} \sim N(60000, \sigma_{inc}^2),\ \sigma_{inc} \sim N^+(0, 1)$$

$$\mu_{age} \sim N(30, \sigma_{age}^2),\ \sigma_{age} \sim N^+(0, 1)$$

$$\beta[0:3] \sim N(0,1)$$

Where

- y_i is 1, if the transaction was online; 0, if transaction was in any other way;
- $x_{i,1}$ is amount of transaction
- $x_{i,2}$ is type of purchase
- $x_{i,3}$ is an indicator variable for the time of purchase which assumes the following values:
  - 1, if the purchase occurred *during the social distancing*;
  - 0, otherwise
- α_m^{inc} is the random effect of the income level, with $m = 1, \ldots, M$;
- α_g^{age} is the random effect of the age group, with $g = 1, \ldots, G$;
- M is the total number of income levels ($M = 4$ in this problem);
- G is the total number of age groups ($G = 4$ in this problem);
- μ_{inc} is the mean income level;
- μ_{age} is the mean age group;
- σ_{inc}^2 is the variance of frequency of income level;
- σ_{age}^2 is the variance of frequency of age group;

This model would be able to answer questions like the proportion of purchases from a specific `age-group` at level g by looking at the parameter α_g^{age} and to questions related to `income` at level m by analyzing the parameter α_m^{inc}

(b) Based on the posterior samples on your model parameters, how would you estimate the expected change in the proportion of online purchases for people aged 18-29 in the first income bracket and construct a 95% credible interval? How would you predict the change in the proportion of online purchases and construct a 95% prediction interval for a group of 30 people aged 18-29 in the first income bracket? Show working with formulas and pseudo code, where appropriate.

{*Answer.*}

The expected change of *proportion online* for people from `age-group` 18-29 ($g = 1$) and first `income` bracket ($m = 1$) can be evaluated by analyzing the MCMC samples of parameters $\alpha_{[1]}^{age}$ and $\alpha_{[1]}^{inc}$.

To calculate credible 95% intervals we will use the estimates of

- Step 1 - Obtain $\hat{beta}_0, \hat{beta}_1, \hat{beta}_2, \hat{beta}_3, \alpha_{[1]}^{inc}, \alpha_{[1]}^{age}$;
- Step 2 - Calculate quantiles 0.025 and 0.975 using the estimates parameters calculated by the model, i.e., by selecting all people which matches the search criteria (i.e., `age-group` 18-29 & `income`="less $30,000") and apply the proper parameter estimates which can be represented by:

$$logit^{-1}(\hat{\pi}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,3} + \hat{\alpha}_{1[i]}^{inc} + \hat{\alpha}_{1[i]}^{age} \tag{1}$$

- Step 3 - Analyse the behavior of specific parameters of interest (in our case $\hat{\beta}_3$ which provides evidence of the change to online transactions. By analyzing the *signal* (positive/negative) to understand if we have increase/reduction in the proportion and estimate the proportion using its *value* which will provide information about the magnitude of change. In this context, its important note that we need to use the function $logit^{-1}$.

In order to estimate the the C.I. we would use the function `Bernoulli_logit` from STAN and calculate the quantiles generated by $logit^{-1}$ with the desired level of confidence, in this case, $[0.025, 0.975]$.

Using the same approach we would estimate the proportion of online purchases for a specific group of 30 people, `age-group` 18-29 ($g = 1$) & `income`="less $30,000" ($m = 1$), which is given by $\hat{\alpha}_1^{inc}$ and $\hat{\alpha}_1^{age}$.

We would select the group of interest by matching the people in that criteria and by selecting a random group within it, by use of something like the pseudo-code below:

- L0 := sample(which(db$ageG==1 & db$inc==1),30,replace=FALSE)
- L1 := which(db[L0,]$yrep==1)[1]
- prop:=length(L1)/length(L0)

The estimation of C.I. would be in the same way already described above, but now with the sample selected.

Now let's focus on the period since social distancing started (i.e. March 16-April 6). In addition, to simplify things, let's consider the effect of age group only. Define y_i to be the number of online purchases for individual i, and n_i to be the total number of purchases for individual i.

KOHO would like to use their data to estimate the proportion of purchases made online for the whole of Toronto. The types of people who use KOHO are not really representative of the broader Toronto; in general they tend to be relatively young. However, all is not lost: from the last census we have population counts by age group in Toronto. So we could post-stratify to get a more representative estimate, i.e.

$$\hat{\pi}^{ps} = \frac{\sum_G \hat{\pi}_g N_g}{N}$$

where $\hat{\pi}^{ps}$ is the estimated proportion of purchases made online, g refers to a particular age group (i.e. there are $G = 4$ total age groups), $\hat{\pi}_g$ is the estimated proportion for a particular group, N_g is the number of people in a particular age group based on the census, and N is the total population based on the census.

(c) Assume that differences in the propensity to make online purchases is fully captured by differences in age. Let π^* be the true proportion of purchases that are made online in Toronto during March 16-April 6. If $\hat{\pi}_g$ is taken to be the observed proportions from the data (i.e. the MLE), show that $\hat{\pi}^{ps}$ is an unbiased estimator of π^*.

{*Answer.*}

Considering the period of social distancing from March 16th to April 6th and the effect of α^{age}, we have:

- y_i: number of online purchases for the individual i;
- n_i: total number of purchases of individual i;
- $\hat{\pi}^{PS} = \frac{\sum_G \hat{\pi_g} N_g}{N}$ where
  - $\hat{\pi}^{PS}$: estimated proportion of purchases online;
  - g: any particular age-group;
  - $\hat{\pi}_g$: estimated proportion of online purchases for a particular group g;
  - N_g: number of people of a particular group given by census;
- π^*: true proportion of purchases online in Toronto.

We need then to show $E(\hat{\pi}^{PS}) = \pi^*$, so let's remember that $\hat{\pi}_g$ is the MLE for π^*, then $\hat{\pi}_g \sim Bin(N_g, \pi^*)$, then:

[1] *yrep* represents the expected people who would buy online calculated by the model

$$\begin{aligned} E(\hat{\pi}^{PS}) &= E\Big(\frac{\sum_G \hat{\pi_g} N_g}{N}\Big) \\ &= \frac{1}{N}\sum_G N_g E\big(\hat{\pi_g}\big) \end{aligned}$$

Please note that $\hat{\pi}_g \sim Bin(N_g, \pi^*)$, then

$$\begin{aligned} E(\hat{\pi}^{PS}) &= \frac{1}{N}\sum_G N_g \frac{\pi^*}{N_g} \\ &= \pi^* \\ \implies E(\hat{\pi}^{PS}) &= \pi^* \end{aligned} \tag{2}$$

Then $\hat{\pi}^{PS}$ is an unbiased estimator for π^*.

(d) Now instead of using raw proportions in the data, you estimate $\hat{\pi}_g^{mr}$ g with a hierarchical model, allowing for a varying intercept α_g by age group, with α_g modeled hierarchically as a draw from a Normal distribution with mean μ_α and variance σ_α^2.

The estimate of the Toronto proportion is:

$$\hat{\pi}^{mrp} = \frac{\sum_G \hat{\pi}_g^{mr} N_g}{N}$$

Find an expression for $\mathrm{E}(\hat{\pi}^{mrp}|\mathbf{y})$. You may assume that the n's are large enough such that the Normal approximation to the Binomial is fine.

{*Answer.*}

Considering that the MLE for $\hat{\pi}_g^{mr}$ is given by

$$\hat{\pi}_g^{mr} = \frac{\sum_{i=1}^{n_g} y_{g[i]}}{n_g} \tag{3}$$

and

$$\sum_{i=1}^{n_g} y_{g[i]} = S_g \sim Bin(n_g, \pi_g^*), \text{ with } g = 1, \ldots, 4 \tag{4}$$

then

$$\begin{aligned} E(\hat{\pi}_g^{mr}|\mathbf{y}) &= E\big(\frac{S_g}{n_g}|\mathbf{y}\big) \\ &= \frac{1}{n_g} E(S_g|\mathbf{y}) \\ \implies E(\hat{\pi}_g^{mr}|\mathbf{y}) &= \frac{1}{n_g} E(S_g|\mathbf{y}) \end{aligned} \tag{5}$$

For n_g larger, by CLT we have that

$$Z = \frac{S_g - n_g\pi_g^*}{\sqrt{n_g\pi_g^*(1-\pi_g^*)}} \sim N(0,1) \tag{6}$$

Note that (5) can be rewritten in the following way:

$$\begin{aligned}
E(\hat{\pi}_g^{mr}|\mathbf{y}) &= \frac{1}{n_g}E(S_g|\mathbf{y}) \\
&= \frac{1}{n_g}E\Big(\sqrt{n_g\pi_g^*(1-\pi_g^*)}Z + n_g\pi_g^*\Big|\mathbf{y}\Big) \\
&= \frac{1}{n_g}\left[\sqrt{n_g\pi_g^*(1-\pi_g^*)}E(Z|\mathbf{y}) + n_g\pi_g^*\right] \\
&= \frac{1}{n_g}\left[0 + n_g\pi_g^*\right] \\
&= \pi_g^*
\end{aligned}$$

$$\implies E(\hat{\pi}_g^{mr}|\mathbf{y}) = \pi_g^*. \tag{7}$$

(e) Calculate the bias of $\hat{\pi}^{mrp}$ given the true proportion π^*. Given it's greater than zero, why would we prefer this estimator over the unbiased $\hat{\pi}^{ps}$? Discuss.

{*Answer.*}

Calculating the Bias of $\hat{\pi}_g^{mr}$ we have the following:

$$\begin{aligned}
Bias(\hat{\pi}_g^{mr}|\pi^*) &= E(\hat{\pi}_g^{mr}|\pi^*) - \pi^* \\
&= E\left(\frac{\sum_G \hat{\pi}_g^{mr}N_g}{N}\middle|\pi^*\right) \\
&= \frac{1}{N}\sum_G N_g E\Big(\hat{\pi}_g^{mr}\Big|\pi^*\Big) \\
&= \frac{1}{N}\sum_G N_g E\Big(\frac{S_g}{N_g}\Big|\pi^*\Big) \\
&= \frac{1}{N}\sum_G E\Big(S_g\Big|\pi^*\Big) \\
&= \frac{1}{N}\sum_G N_g\pi_g > 0
\end{aligned}$$

In this case, $\hat{\pi}^{PS}$ would be preferred, even if biased, because $Var(\hat{\pi}^{PS}) < Var(\hat{\pi}^{mr})$.

2 - Maternal Mortality

This question relates to estimating the maternal mortality for countries worldwide. A maternal death is defined by the World Health Organization as "the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes". The indicator we are interested in is the (non-AIDS) maternal mortality ratio (MMR) which is defined as the number of non-AIDS maternal deaths divided by the number of live births.

In the data folder of the class repo there are two files relevant to this question. `mmr_data` contains information on, for a range of countries over a range of years:

- Observations of the proportion of non-AIDS deaths that are maternal (PM^{NA});
- Data source, most commonly from Vital Registration systems (VR);
- The Gross Domestic Product (GDP);
- The General Fertility Rate (GFR);
- The average number of skilled attendants at birth (SAB)
- The geographical region of the country
- The total number of women, births, deaths to women of reproductive age (WRA), and the estimated proportion of all WRA deaths that are due to HIV/AIDS

The `mmr_data` file will be used for fitting. Note that data on PMNA is not available for every country.

The `mmr_pred` file contains information on GDP, GFR, SAB, total number of births, deaths and women, and proportion of deaths that are due to HIV/AIDS, for every country at different time points (every five years from mid 1985 to mid 2015). Information in this file is used for producing estimates of MMR for countries without data, and for producing estimates centered at a particular time point.

Consider the following model

$$y_i|\eta_{c[i]}^{country}, \eta_{r[i]}^{region} \sim N(\beta_0 + \eta_{c[i]}^{country} + \eta_{r[i]}^{region} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}, \sigma_y^2)$$

$$\eta_c^{country} \sim N\left(0, \left(\sigma_\eta^{country}\right)^2\right), \text{ for } c = 1, 2, \ldots, C$$

$$\eta_r^{region} \sim N\left(0, \left(\sigma_\eta^{region}\right)^2\right), \text{ for } r = 1, 2, \ldots, R$$

where

- y_i is the ith observed log PM^{NA} in a country $c[i]$ in region $r[i]$;
- C is the total number of countries and R is the total number of regions;
- $x_{i,1}$ is log(GDP)
- $x_{i,2}$ is log(GFR)
- $x_{i,3}$ is SAB

(a) Turn this model into a Bayesian model by specifying appropriate prior distributions for the hyper-parameters and fit the Bayesian model in Stan. Report the full model specification as well as providing the Stan model code.

Hint: I would recommend indexing countries and regions, and calculating C and R based on the full set of countries contained in `mmr_pred`, rather than the subset contained in `mmr_data`. This will mean you will automatically get estimates for η for every country and region, even the missing ones, which will help later on.

E.g. to get full list of country iso codes and regions, could do something like:

```r
country_region_list <- mmr_pred %>%
  group_by(iso) %>%
  slice(1) %>%
  arrange(iso) %>%
  select(iso, region)

iso.c <- country_region_list$iso # the iso country of each country
C <- length(iso.c) # number of countries

region.c <- country_region_list$region # the region that country c belongs to (name)
regions <- unique(region.c) # a list of all unique regions
R <- length(regions) # number of regions

# the region index that country c belongs to
r.c <- as.numeric(factor(region.c, levels = regions))
```

Then to get the relevant indexes for each observation i:

```r
c.i <- as.numeric(factor(mmr_data$iso, levels = iso.c))
r.i <- r.c[c.i] # the region of the ith observation
```

{*Answer.*}

Following the proposed model stated by the question, the Bayesian model implemented is as follows:

$$y_i|\eta_{c[i]}^{country}, \eta_{r[i]}^{region} = \beta_0 + \eta_{c[i]}^{country} + \eta_{r[i]}^{region} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}, \sigma_y^2)$$

with the following priors:

$$\begin{gathered}
\beta_i \sim N\big(0,1\big) \text{ for } i = 0, \ldots, 3 \\
\sigma_y \sim N\big(0,1\big) \\
\eta_c^{country} \sim N\left(0, \left(\sigma_\eta^{country}\right)^2\right) \\
\eta_r^{region} \sim N\left(0, \left(\sigma_\eta^{region}\right)^2\right) \\
\sigma_\eta^{country} \sim N\big(0,1\big) \\
\sigma_\eta^{region} \sim N\big(0,1\big)
\end{gathered}$$

where

- y_i is the ith observed log PM^{NA} in a country $c[i]$ in region $r[i]$;
- C is the total number of countries and R is the total number of regions;
- $x_{i,1}$ is log(GDP)
- $x_{i,2}$ is log(GFR)
- $x_{i,3}$ is SAB

The STAN-model was implemented as follows:

```
data {
  int<lower=1> N;                    // number of observations
  int<lower=1> C;                    // Number of countries
  int<lower=1> R;                    // Number of regions
  vector[N] log_gdp;                 // log of GDP of country
  vector[N] log_gfr;                 // log of GFR of country
```

```
  vector[N] sab;                          // SAB of that country
  int<lower=1> country[N];                // country of observation
  int<lower=1> region[N];                 // region of observation
  vector[N] y;                            // log of PMNA
}

parameters {
  real eta_country[C];            // eta country
  real eta_region[R];             // eta region
  real beta[4];                   // beta coefficients
  real<lower=0> sigma;            // sigma of log_pmna
  real<lower=0> sigma_country;    // sigma eta country
  real<lower=0> sigma_region;     // sigma eta region
}

model {
  vector[N] y_hat;

  beta ~ normal(0, 1);
  sigma ~ normal (0,1);
  sigma_country ~ normal (0, 1);
  sigma_region ~ normal (0, 1);
  eta_country ~ normal (0, sigma_country);
  eta_region~ normal (0, sigma_region);

  for (i in 1:N)
      y_hat[i] = beta[1] + eta_country[country[i]] + eta_region[region[i]] +
                           beta[2]*log_gdp[i] + beta[3]*log_gfr[i] +  beta[4]*sab[i];

  y ~ normal(y_hat, sigma);

}

generated quantities {
  vector[N] y_rep;     // replications from posterior predictive dist
  vector[N] log_lik;   // pointwise log-likelihood for LOO

  for (n in 1:N) {
    y_rep[n] = normal_rng(beta[1] +
                          eta_country[country[n]] +
                          eta_region[region[n]] +
                          beta[2]*log_gdp[n] + beta[3]*log_gfr[n] +
                          beta[4]*sab[n], sigma);

    log_lik[n] = normal_lpdf( y[n] | beta[1] +
                                eta_country[country[n]] +
                                eta_region[region[n]] +
                                beta[2]*log_gdp[n] + beta[3]*log_gfr[n] +
                                beta[4]*sab[n], sigma);
  }

}
```

(b) Check the trace plots and effective sample size to check convergence and mixing. Summarize your findings

using a few example trace plots and effective sample sizes.

{*Answer.*}

The diagnostics for the adjusted model seems to be a good fit:

- the chains are well mixed as we can see in `traceplot`;
- `RHat` is distributed around 1.0;
- estimators seems to converge to the true expectation as we can see through `pairs` plots.

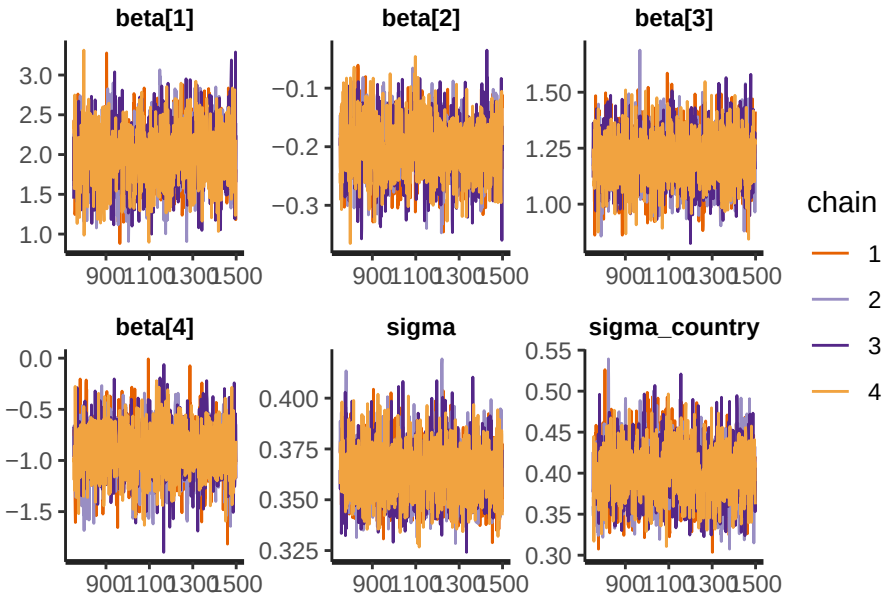


Figure 1: Diagnostics from Model PMNA - Traceplots

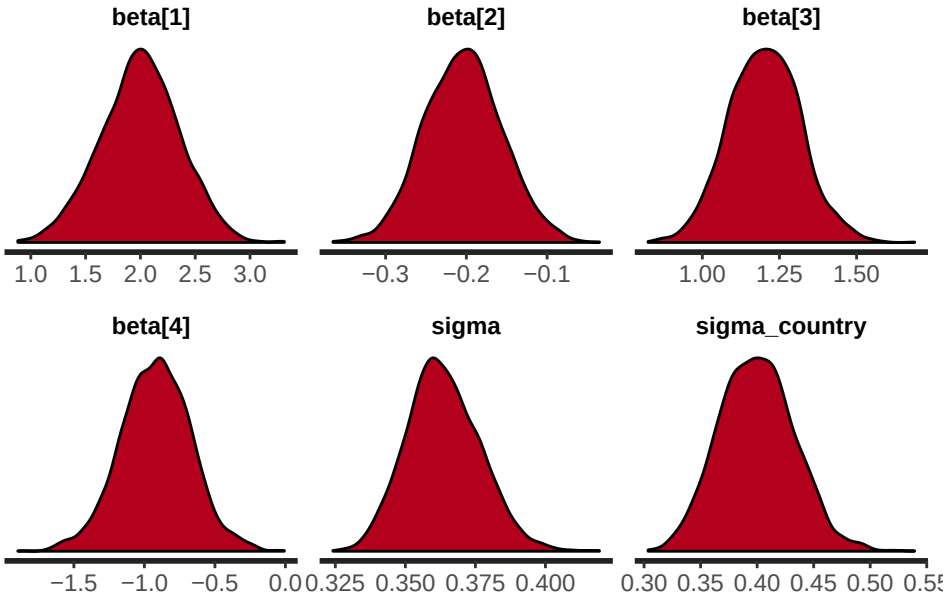


Figure 2: Diagnostics from Model PMNA - Parameters Densities

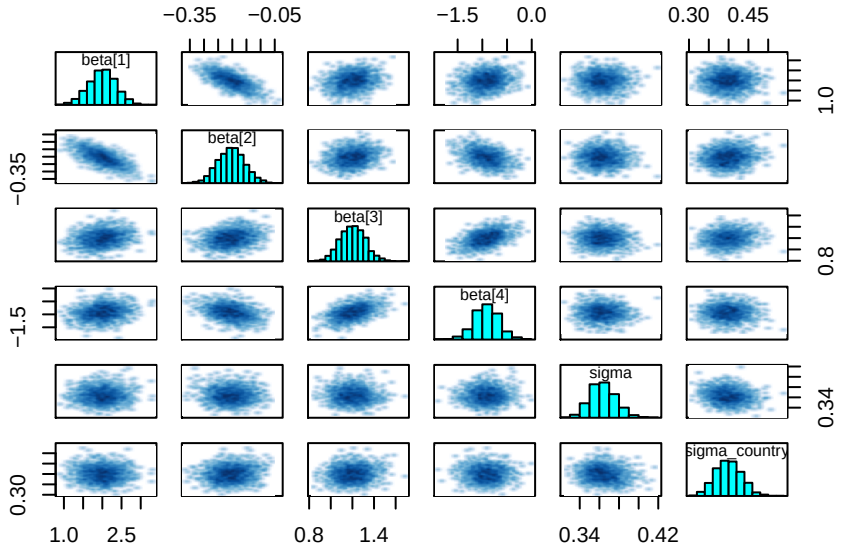


Figure 3: Diagnostics from Model PMNA - Pair-plot

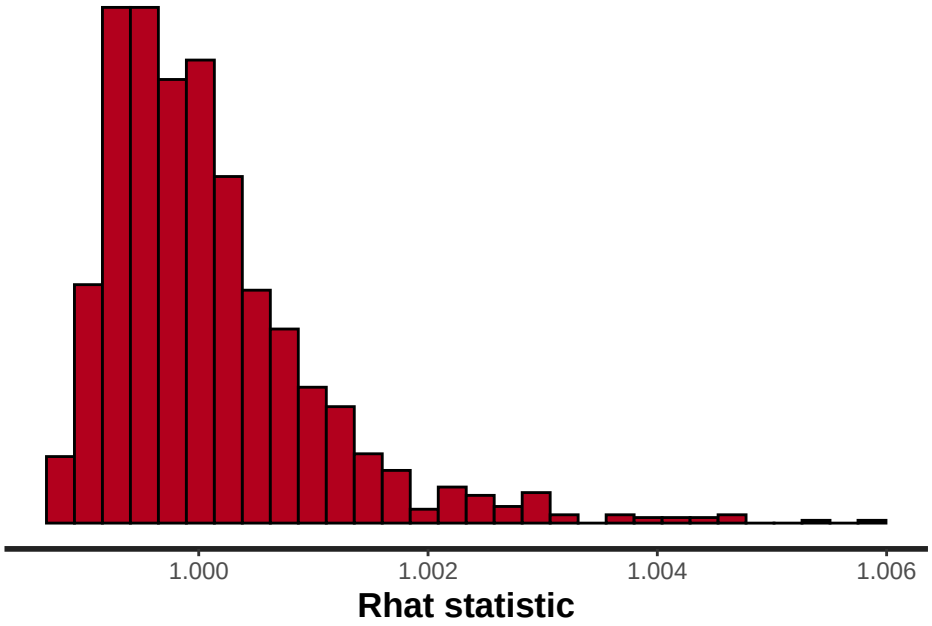
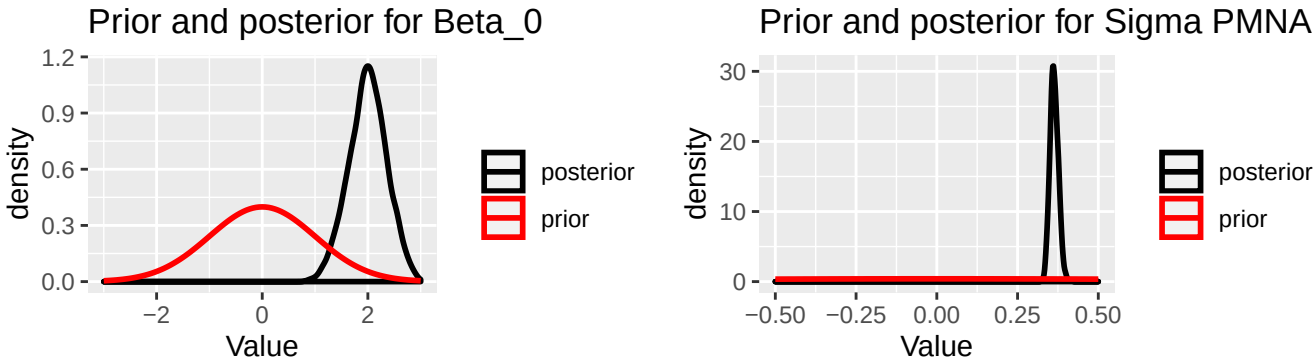


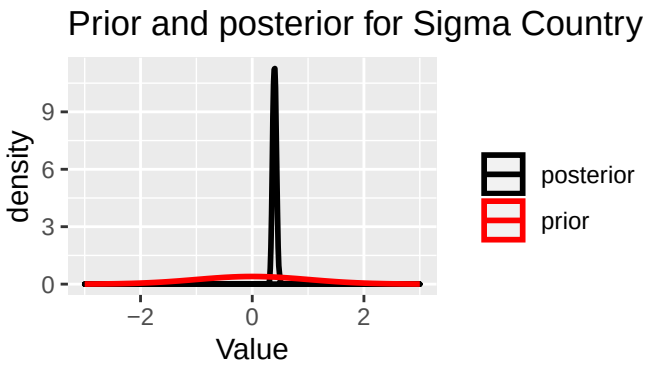
Figure 4: Diagnostics from Model PMNA - RHat

(c) Plot (samples of the) prior and posterior distributions for $\beta_0, \sigma_y, \sigma_y^{country}$ and σ_η^{region}. Interpret the estimates of β_1 and β_3.

{*Answer.*}

Plotting the prior and posterior densities for the parameters we obtained the following:





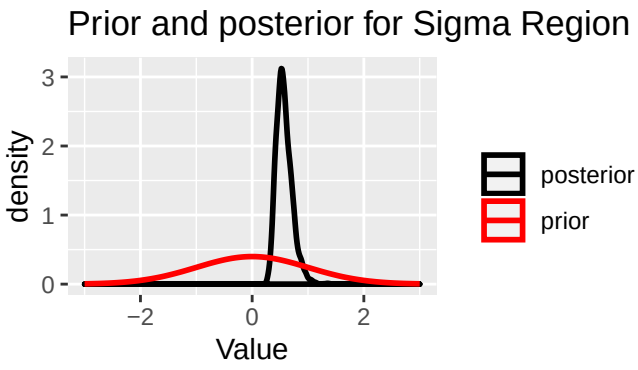


Table 1: Coefficients for Beta

	mean	se_mean	sd	Rhat
beta[1]	1.9997696	0.0133232	0.3598012	1.004375
beta[2]	-0.2036981	0.0017917	0.0479330	1.003131
beta[3]	1.2023199	0.0029838	0.1158932	1.000362
beta[4]	-0.9196816	0.0078105	0.2492078	1.003788

From model adjustment the estimates we obtained are $\beta_1 = -0.2037$ (for log(GDP)) and $\beta_3 = -0.9197$ (for SAB). This means that increases of `GDP` and `SAB` will represent *decreases* in `PMNA` which makes sense. For instance, a 10% increase in `GDP` will represent a 2.07% decrease on `log(PMNA)`. In the same way, if we have an increase in the number of skilled attendant at birth (SBA) of 10%, this will represent a decrease of equal amount at `log(PMNA)`.

(d) Use the MCMC samples to construct 95% credible intervals for the PMNA for 5-year periods from 1985.5 to 2015.5 for one country with data and one country without any observed PM^{NA} values. Provide point estimates and CIs in a table and a nice plot. Add the observed data to the plot as well (for the country that has it).

{*Answer.*}

First we need to select both countries by comparing the input databases, i.e., find one country that is present on both **Training-DB** (`mmr_data`) and **Test-DB** (`mmr_pred`), and another one that is present only on training database.

I selected two countries at random to elaborate the C.I.s.

```
## # A tibble: 0 x 2
## # Groups:   iso [1]
## # ... with 2 variables: iso <fct>, region <fct>
```

For **Jamaica** we have the following graph.

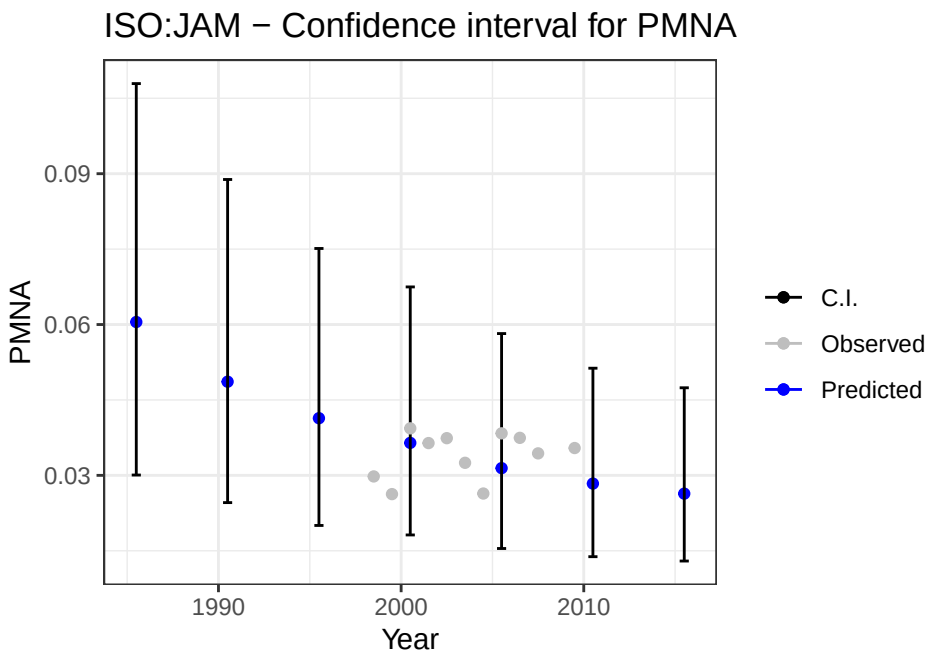


Figure 5: 95% Confidence interval for PMNA - Both DB

For **Qatar** we have the following graph.

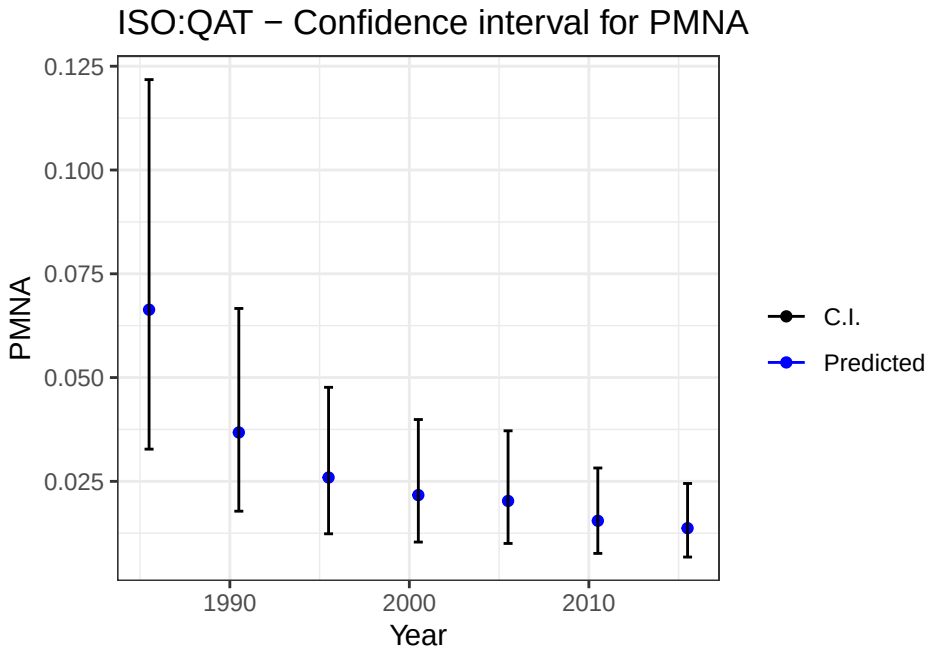


Figure 6: 95% Confidence interval for PMNA - Only on Pred-DB

(e) The non-AIDS MMR is given by:

$$\begin{aligned} MMR^{NA} &= \frac{\text{\# Non-AIDS maternal deaths}}{\text{\# of Births}} \\ &= \frac{\text{\# Non-AIDS maternal deaths}}{\text{\# Non-AIDS deaths}} \times \frac{\text{\# Non-AIDS deaths}}{\text{\# of Births}} \\ &= PM^{NA} \times \frac{\text{\# Deaths * (1 - prop AIDS)}}{Births} \end{aligned}$$

where deaths and births are to all women of reproductive age in the country-period of interest.

Use this formula, your answers from d) and the data in `mmr_pred` to obtain point estimates and CIs for the non-AIDS MMR for the two countries you chose in (d) in the year 2010.5.

{*Answer.*}

Now we will estimate the non-AIDS MMR for both countries from previous item, i.e. **Jamaica** and **Qatar**.

The CI's for 2010 for both countries are:

Table 2: C.I. for MMR in 2010

Country	MMR	CIUpper	CILower
Jamaica	0.0007557	0.0013672	0.0003683
Qatar	0.0001448	0.0002620	0.0000706

(f) In the model used so far, we assume that error variance σ_y^2 is the same for all observations but this is probably not a very realistic assumption. Let's explore if the model fit changes if we would estimate two variance parameters: one for VR data (denoted by σ_{VR}^2) and one for non-VR data (denoted by σ_{non-VR}^2). Write out the model specification for this extended model, give the Stan model code, and fit the model. Show priors and posteriors for σ_{VR}^2 and σ_{non-VR}^2 and construct a plot with data for a country with VR data, with point estimates and CIs from the models with and without equal variance.

{*Answer.*}

In order to consider that the variable of interest y_i can have different variances depending if the source is from *VR* or *non-VR* we will add 01 (one) additional component γ_{VR}which stands for if the data is VR-sources with respective variance denoted by σ_{VR}^2 .

In this sense, our new *enhanced* model becomes:

$$y_i|\eta_{c[i]}^{country}, \eta_{r[i]}^{region} \sim N(\beta_0 + \eta_{c[i]}^{country} + \eta_{r[i]}^{region} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}, \sigma_{VR}^2 x_{i,4} + \sigma_{non-VR}^2 (!x_{i,4}))$$

$$\eta_c^{country} \sim N\left(0, \left(\sigma_\eta^{country}\right)^2\right), \text{ for } c = 1, 2, \ldots, C$$

$$\eta_r^{region} \sim N\left(0, \left(\sigma_\eta^{region}\right)^2\right), \text{ for } r = 1, 2, \ldots, R$$

where

- y_i is the ith observed log PM^{NA} in a country $c[i]$ in region $r[i]$;
- C is the total number of countries and R is the total number of regions;
- σ_{VR}^2 is the variance for VR-type data;
- σ_{non-VR}^2 is the variance for nonVR-type data;
- $x_{i,1}$ is log(GDP)

- $x_{i,2}$ is log(GFR)
- $x_{i,3}$ is SAB
- $x_{i,4}$ is an indicator variable standing which assumes 1 if the data is VR, 0 otherwise;

The new STAN-model implemented is as follows:

```
data {
  int<lower=1> N;                       // number of observations
  int<lower=1> C;                       // Number of countries
  int<lower=1> R;                       // Number of regions
  vector[N] log_gdp;                    // log of GDP of country
  vector[N] log_gfr;                    // log of GFR of country
  vector[N] sab;                        // SAB of that country
  int<lower=0, upper=1> vr[N];          // binary variable for VR-data
  int<lower=1> country[N];              // country of observation
  int<lower=1> region[N];               // region of observation
  vector[N] y;                          // log of PMNA
}

parameters {
  real eta_country[C];            // eta country
  real eta_region[R];             // eta region
  real beta[4];                   // beta coefficients
  real<lower=0> sigma_vr;         // sigma of VR
  real<lower=0> sigma_nvr;        // sigma of non-VR
  real<lower=0> sigma_country;    // sigma eta country
  real<lower=0> sigma_region;     // sigma eta region
}

model {
  vector[N] y_hat;
  vector[N] sigma;

  beta ~ normal (0, 1);
  sigma_vr ~ normal (0, 1);
  sigma_nvr ~ normal (0, 1);
  sigma_country ~ normal (0, 1);
  sigma_region ~ normal (0, 1);
  eta_country ~ normal (0, sigma_country);
  eta_region~ normal (0, sigma_region);

  for (i in 1:N) {
      y_hat[i] = beta[1] + eta_country[country[i]] + eta_region[region[i]] +
                           beta[2]*log_gdp[i] + beta[3]*log_gfr[i] + beta[4]*sab[i];
      sigma[i] = sigma_vr*vr[i] + sigma_nvr*(!vr[i]);
  }
  y ~ normal(y_hat, sigma);

}
```

Running the new model...

We can see that the model converged and the estimates of the parameters seems reliable.

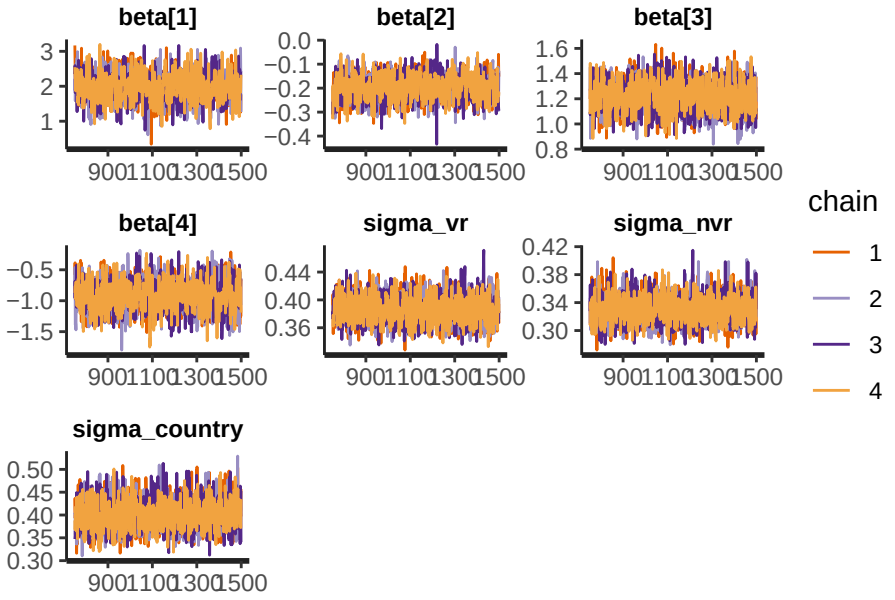


Figure 7: Diagnostics from Enhanced Model PMNA - Traceplots

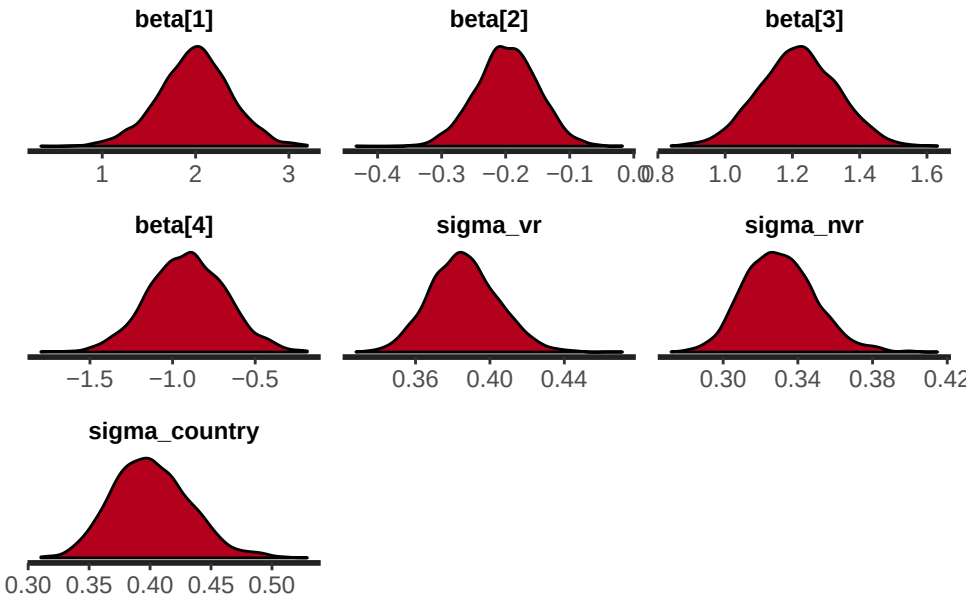


Figure 8: Diagnostics from Enhanced Model PMNA - Parameters Densities

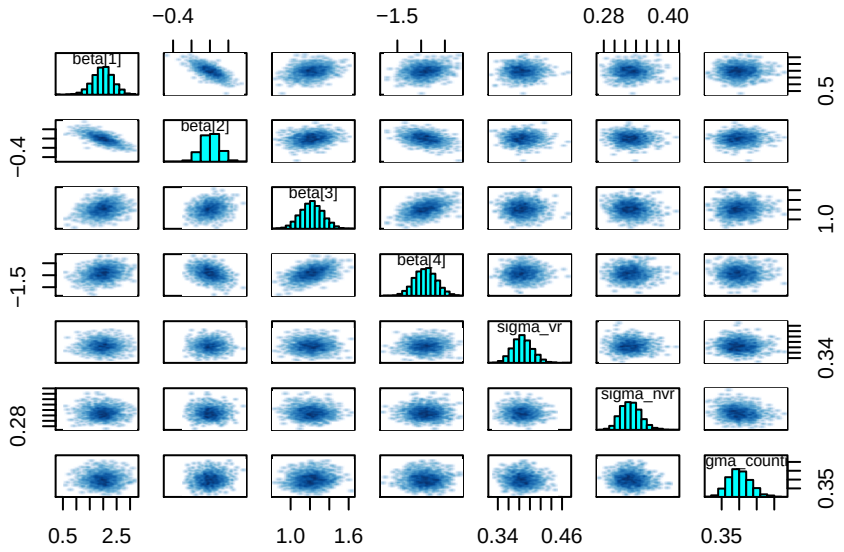


Figure 9: Diagnostics from Enhanced Model PMNA - Pair-plots

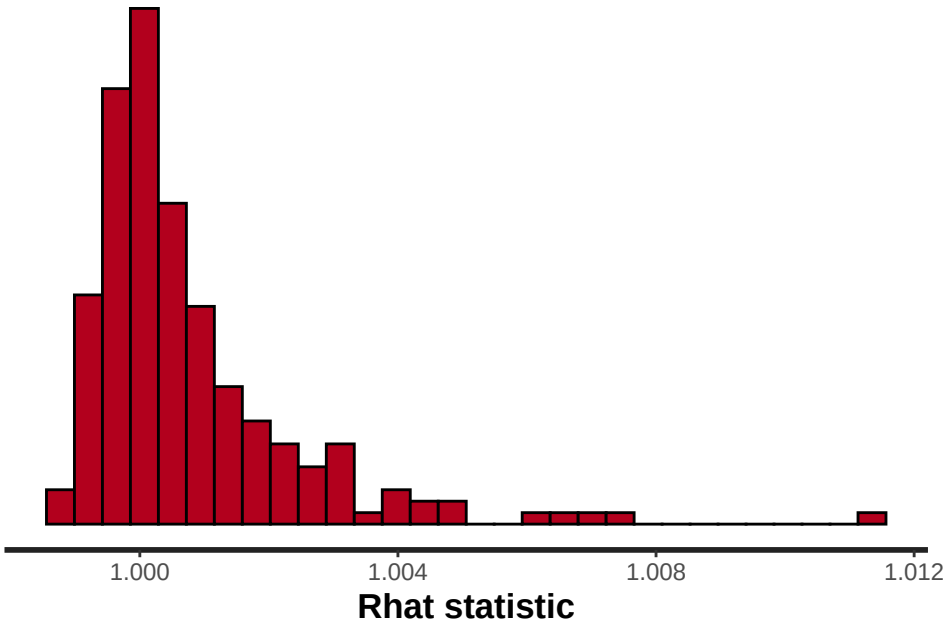
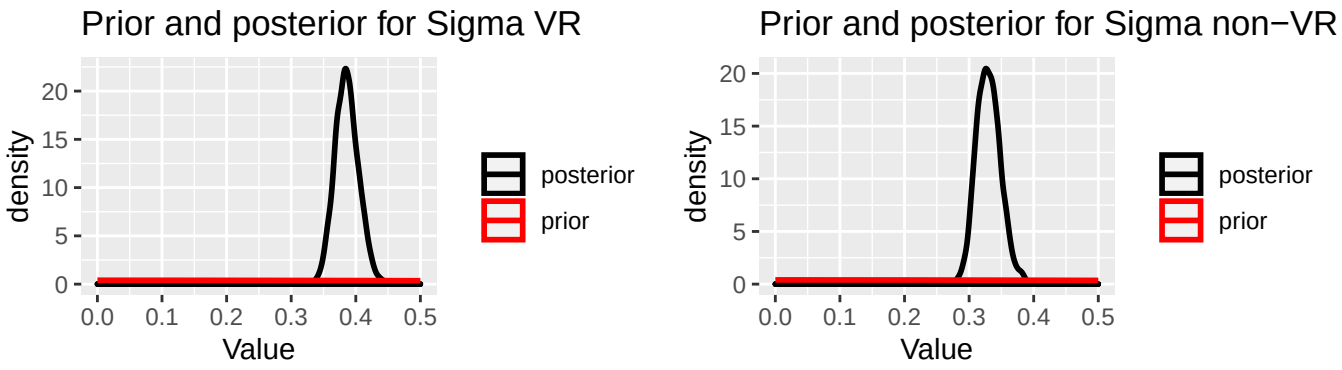


Figure 10: Diagnostics from Enhanced Model PMNA - RHat

Plotting the prior and posterior densities for the parameters of the new model we obtained the following:



Now plotting the C.I.s for **Argentina**[2] which has VR-type data, we have the following:

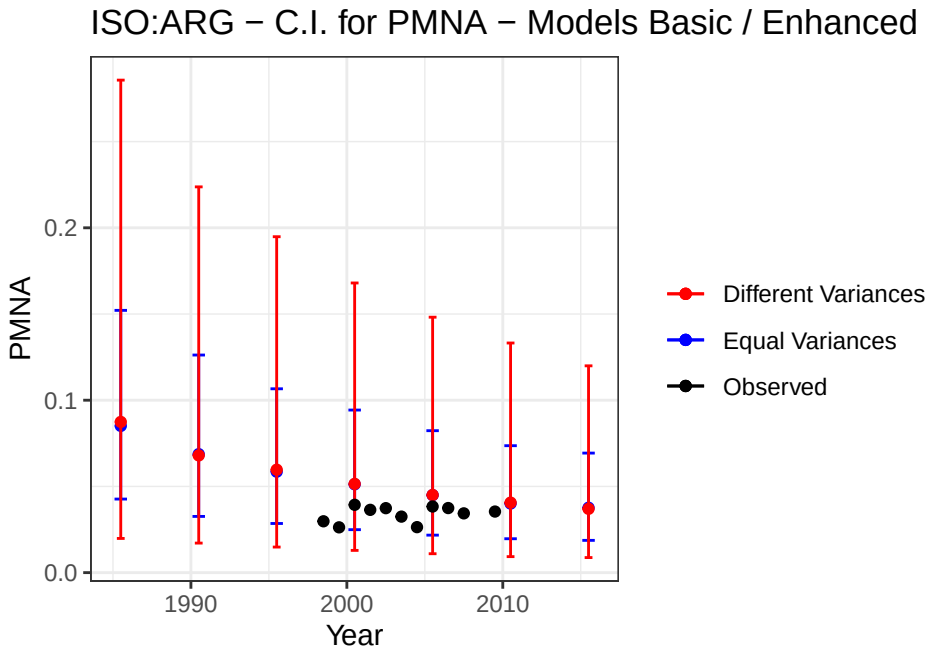


Figure 11: 95% Confidence interval for PMNA - Equal/Different variances

[2]The R-code was reused from previous items with minimum renaming of variables to optimize the development.

3 - Airbnb

In this question you will be exploring what factors are associated with nightly rates of accommodation listed on Airbnb in Toronto. In the data folder of the class repo there is a file called airbnb. This contains variables describing Airbnb listings in Toronto as of 7 December 2019. I downloaded this from the Inside Airbnb website: `http://insideairbnb.com/get-the-data.html`. I restricted the dataset in the repo to only contain a selection of all variables available, but other than that made no changes.

The goal is to model `price` (or `log(price)` might be more appropriate).

(a) Carry out EDA on this data set. Note that this should include checking the data for missing values, data quality, etc, as well as a descriptive analysis of the data (keeping in mind the modeling goal). As a start, you'll notice that the `price` column is a character, which isn't helpful. Here's some code to get it into a number:

```
airbnb <- airbnb %>%
  mutate(price = str_remove(price, "\\$"),
         price = str_remove(price, ","),
         price = as.integer(price)
  )
```

Note that for the next questions it is okay with me to remove some of the troublesome observations, e.g. I removed observations with missing values for variables like `review_scores_rating`. Just make sure you document what you're removing.

{*Answer.*}

Performing a fist overview of data and looking first for general aspects of data.

```
##     host_id            host_since                 host_response_time
##  Min.   :     1565   Min.   :2008-08-08   a few days or more:  321
##  1st Qu.: 20084147   1st Qu.:2014-08-25   N/A               : 6324
##  Median : 63899880   Median :2016-03-23   within a day      : 1584
##  Mean   :100518757   Mean   :2016-03-15   within a few hours: 2542
##  3rd Qu.:170633488   3rd Qu.:2018-01-24   within an hour    :12341
##  Max.   :315115165   Max.   :2019-12-06   NA's              :  285
##                      NA's   :285
##  host_is_superhost host_listings_count host_total_listings_count
##  Mode :logical     Min.   :  0.000     Min.   :  0.000
##  FALSE:16870       1st Qu.:  1.000     1st Qu.:  1.000
##  TRUE :6242        Median :  2.000     Median :  2.000
##  NA's :285         Mean   :  6.142     Mean   :  6.142
##                    3rd Qu.:  4.000     3rd Qu.:  4.000
##                    Max.   :328.000     Max.   :328.000
##                    NA's   :285         NA's   :285
##                          neighbourhood_cleansed           room_type
##  Waterfront Communities-The Island  : 4375      Entire home/apt:15109
##  Niagara                            : 1048      Hotel room     :   91
##  Annex                              :  851      Private room   : 7784
##  Church-Yonge Corridor              :  722      Shared room    :  413
##  Bay Street Corridor                :  686
##  Dovercourt-Wallace Emerson-Junction:  635
##  (Other)                            :15080
##    bathrooms      accommodates       bedrooms       square_feet
##  Min.   :0.000   Min.   : 1.000   Min.   : 0.000   Min.   :    0.0
##  1st Qu.:1.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.:    0.0
```

```
##  Median :1.000   Median : 2.000   Median : 1.000   Median :    0.0
##  Mean   :1.251   Mean   : 3.169   Mean   : 1.322   Mean   :  656.7
##  3rd Qu.:1.000   3rd Qu.: 4.000   3rd Qu.: 2.000   3rd Qu.:  775.0
##  Max.   :8.500   Max.   :16.000   Max.   :15.000   Max.   :16146.0
##  NA's   :10                       NA's   :25       NA's   :23254
##      price         number_of_reviews has_availability review_scores_rating
##  Min.   :    0.0   Min.   :  0.00    Mode:logical     Min.   : 20.0
##  1st Qu.:   65.0   1st Qu.:  1.00    TRUE:23397       1st Qu.: 92.0
##  Median :   99.0   Median :  8.00                     Median : 97.0
##  Mean   :  148.7   Mean   : 27.94                     Mean   : 94.2
##  3rd Qu.:  160.0   3rd Qu.: 31.00                     3rd Qu.:100.0
##  Max.   :13255.0   Max.   :779.00                     Max.   :100.0
##                                                       NA's   :4676
##  review_scores_accuracy review_scores_cleanliness review_scores_checkin
##  Min.   : 2.000         Min.   : 2.000            Min.   : 2.000
##  1st Qu.: 9.000         1st Qu.: 9.000            1st Qu.:10.000
##  Median :10.000         Median :10.000            Median :10.000
##  Mean   : 9.638         Mean   : 9.408            Mean   : 9.721
##  3rd Qu.:10.000         3rd Qu.:10.000            3rd Qu.:10.000
##  Max.   :10.000         Max.   :10.000            Max.   :10.000
##  NA's   :4687           NA's   :4687              NA's   :4689
##  review_scores_communication review_scores_location review_scores_value
##  Min.   : 2.000              Min.   : 2.000         Min.   : 2.00
##  1st Qu.:10.000              1st Qu.:10.000         1st Qu.: 9.00
##  Median :10.000              Median :10.000         Median :10.00
##  Mean   : 9.741              Mean   : 9.717         Mean   : 9.44
##  3rd Qu.:10.000              3rd Qu.:10.000         3rd Qu.:10.00
##  Max.   :10.000              Max.   :10.000         Max.   :10.00
##  NA's   :4686                NA's   :4693           NA's   :4691
##    host_time          log_HLC          log_HTLC        log_price
##  Min.   : 0.3589   Min.   :  -Inf   Min.   :  -Inf   Min.   : -Inf
##  1st Qu.: 2.2240   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:4.174
##  Median : 4.0658   Median :0.6932   Median :0.6932   Median :4.595
##  Mean   : 4.0872   Mean   :  -Inf   Mean   :  -Inf   Mean   : -Inf
##  3rd Qu.: 5.6438   3rd Qu.:1.3863   3rd Qu.:1.3863   3rd Qu.:5.075
##  Max.   :11.6931   Max.   :5.7930   Max.   :5.7930   Max.   :9.492
##  NA's   :285       NA's   :285      NA's   :285
##  room_type_cod   host_resp_cod                                      neighb
##  Min.   :1.000   Length:23397       Waterfront Communities-The Island  : 4375
##  1st Qu.:1.000   Class :character   Niagara                            : 1048
##  Median :1.000   Mode  :character   Annex                              :  851
##  Mean   :1.722                      Church-Yonge Corridor              :  722
##  3rd Qu.:3.000                      Bay Street Corridor                :  686
##  Max.   :4.000                      Dovercourt-Wallace Emerson-Junction:  635
##                                     (Other)                            :15080
##      shost
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2701
##  3rd Qu.:1.0000
##  Max.   :1.0000
##  NA's   :285
```

In order to maintain in the database only relevant data related to reviews from customers, we will remove:

- missing values from variables `review_score_rating`-type which represents no relevant information in related to customer experience;
- missing values from variables related to host information, which doesn't enable to analyze host characteristics that might influence the `price`;
- observations with `price` equals *zero*;
- column `square_feet` which accounts with 99.4% of its observations with missing values
- column `has_availability` as all observations contains `TRUE` representing units with availability.

Then variable of interest is `log(price)` which seems to have a normal distribution, as we can see by the density distribution of observations.

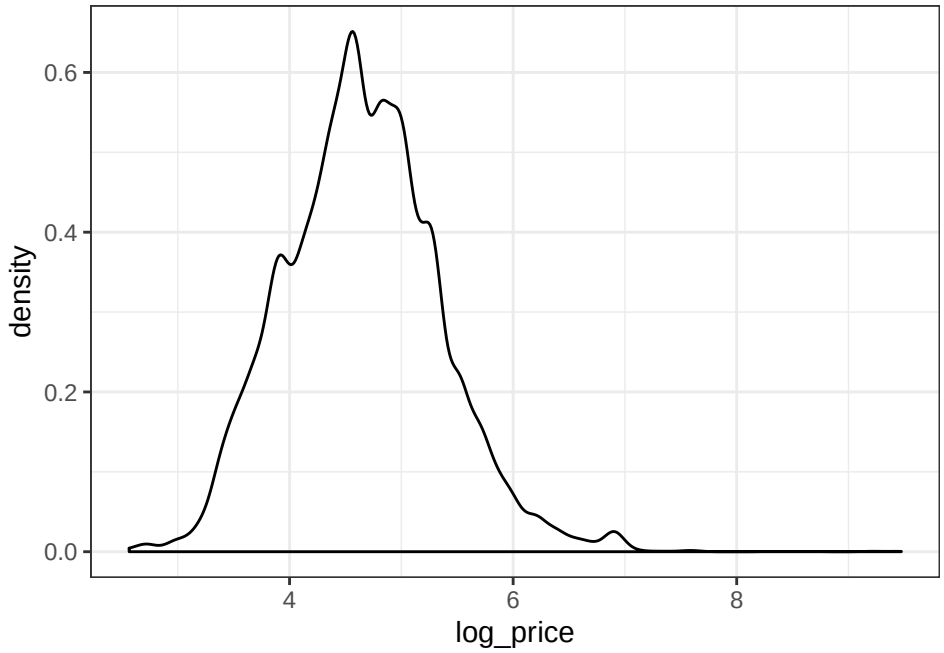


Figure 12: Density of Observed Data - Variable log(price)

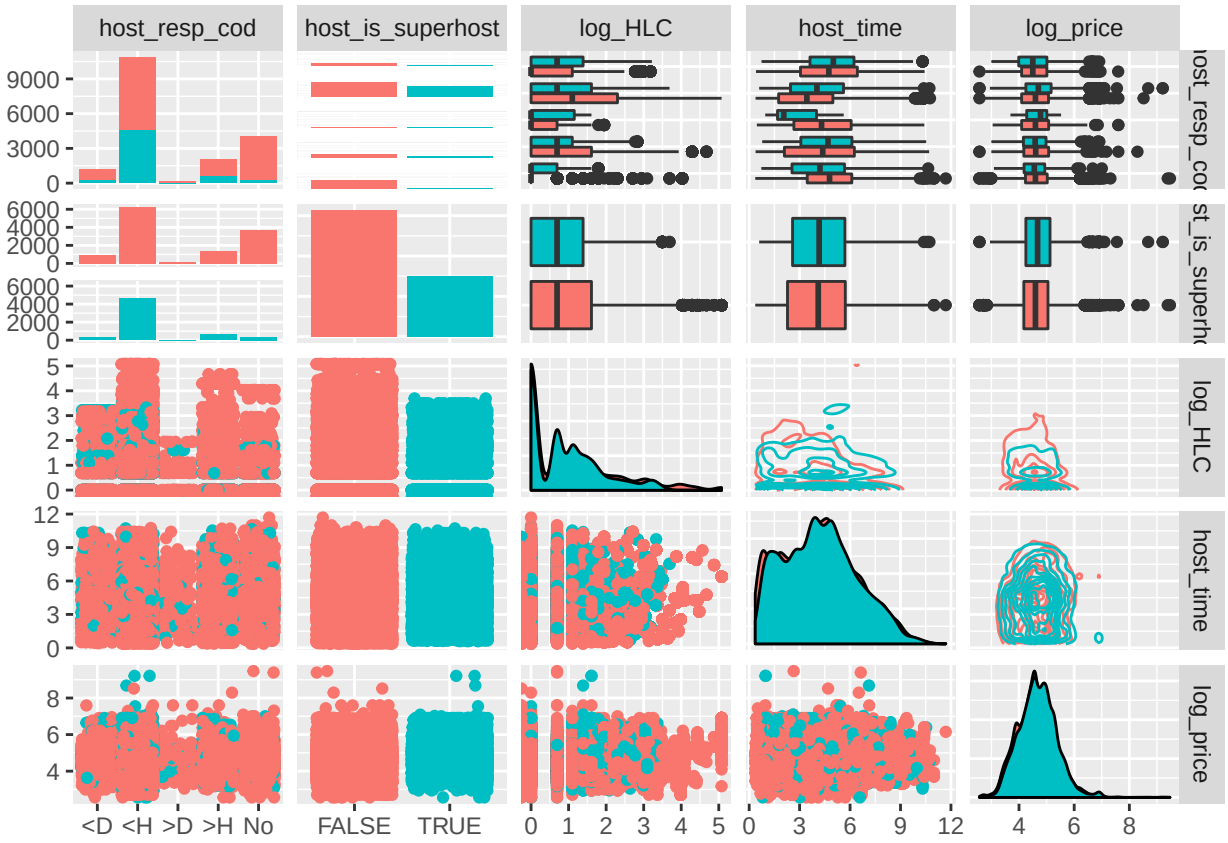


Figure 13: Cross-Correlation between Price & Host

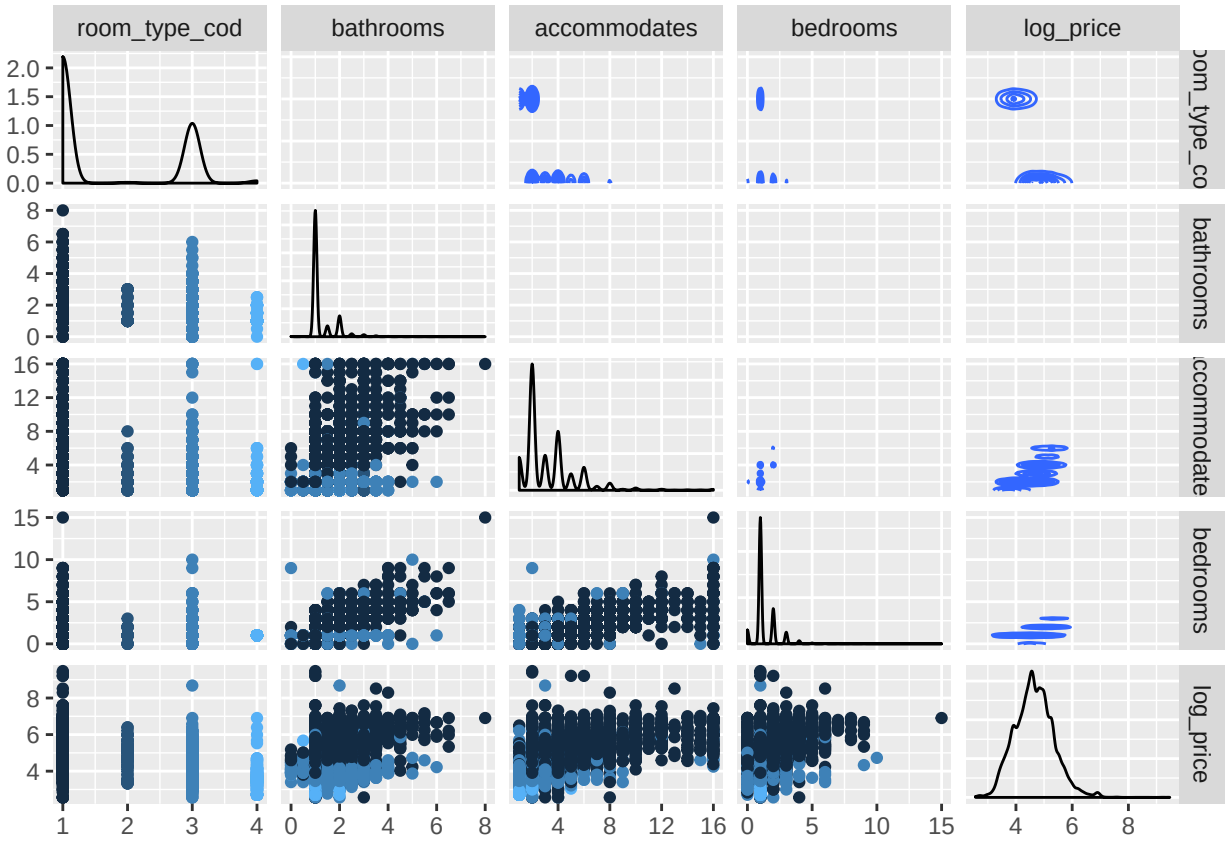


Figure 14: Cross-Correlation between Price & Unit

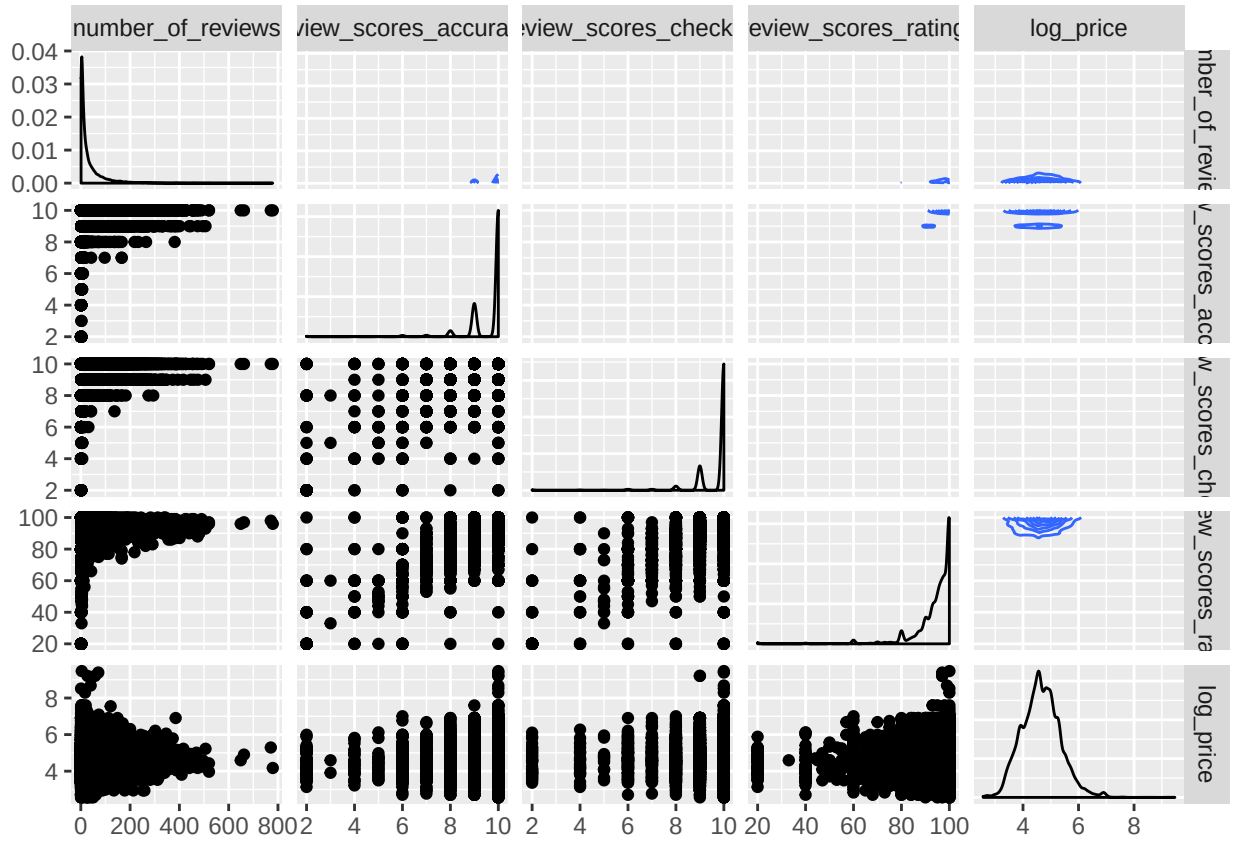


Figure 15: Cross-Correlation between Price & Customer Experience

The heat-maps below will provide additional view of correlation between covariates.

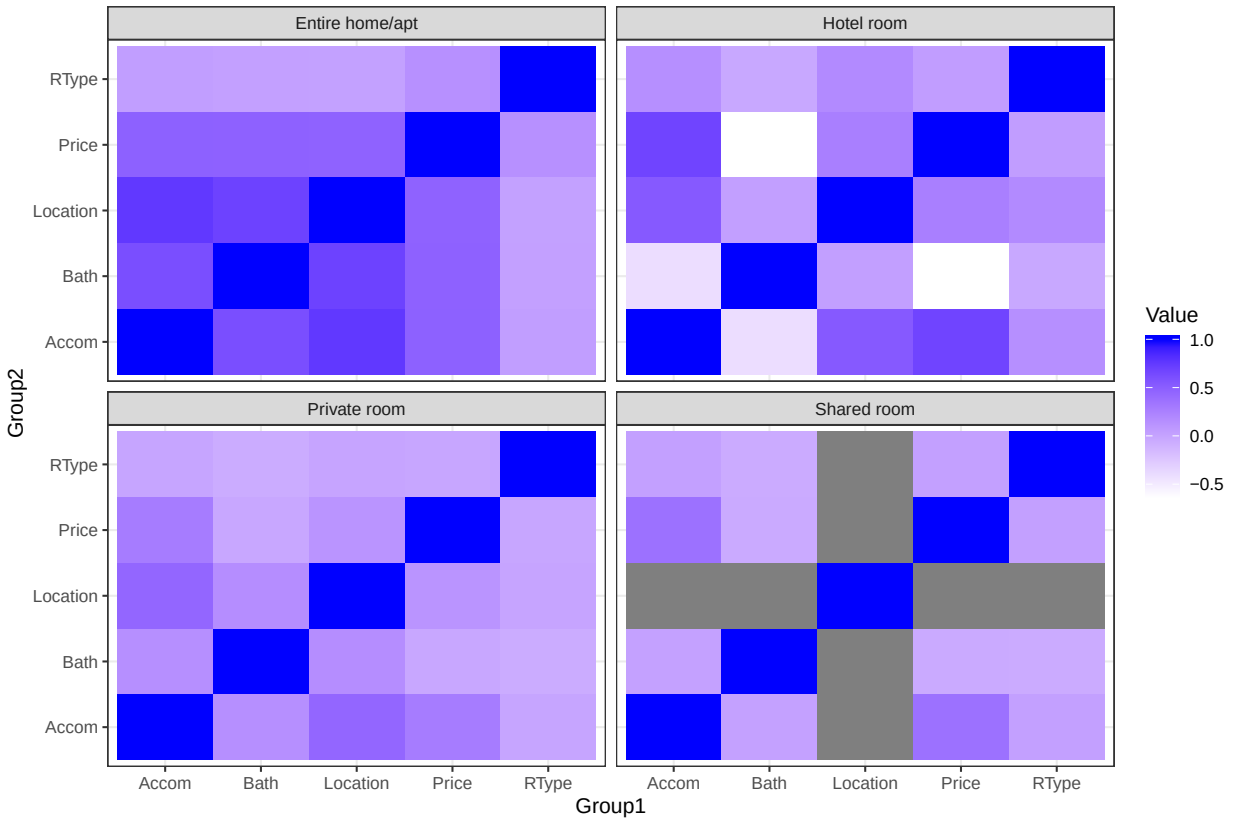


Figure 16: Correlation of variables - Dimension: PLACE

As we can see, there is a different pattern of correlation between variables by `room_type` suggesting this variable as candidate for our target model.

The general conclusion for next questions is to concentrate on some specific variables and then growing the model complexity to fine-tune and get better posterior estimates. From the pairs of variables, we can see the sample distribution of `bedrooms`, `accommodates` and `bathrooms` seems to be more correlated with `price`. The other variables related to **HOST** and `CUSTOMER` might be explored as well, but we will first concentrate our efforts on variables related to variables associated to the unit. Further studies/modelling might be done using different dimensions to include other aspects to the rental marketplace.

(b) Propose two candidate Bayesian models for `log(price)` and fit the two models in Stan (note: you will need to calculate the log-likelihood for the next question). Make sure the model converged and mixing is fine by checking model diagnostics.

{*Answer.*}

In order to model the `price` (which I will be hereon referring to `log(price)`), I considered that this variable *is* (or *maybe*) attached in most cases to 03 main entities: **Place** which stands for the qualifications of the unit, space, location, size etc; **Customer** which reflects the previous experiences with the pair Unit/Host and, in some kind, reflect the demand for that place; and ultimately by **Host** which is the qualification of the host/landlord in offering units to this marketplace.

As we can see from above EDA, the `price` is influenced by variables associated to the unit but possibly with different intensity as we have different shapes of heatmaps per `room_type`. Another dimension that has some hierarchy over the `price` is the location of the unit.

That being said, I will propose 02 models[3]:

Model 1

$$y_i \sim N\Big(\beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,2}, \sigma_y^2\Big)$$

where

- y_i is the ith observed `log(price)`;
- $x_{i,1}$ is number of `bedrooms` of ith observation;
- $x_{i,2}$ is number of `accommodates` of ith observation;
- $x_{i,3}$ is number of `bathrooms` of ith observation;

Model 2

$$y_i \sim N\Big(\beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,2} + \eta_{l[i]}^{Location} + \eta_{r[i]}^{Room}, \sigma_y^2\Big)$$
$$\eta_c^{Location} \sim N\left(0, \left(\sigma_\eta^{Location}\right)^2\right), \text{ for } l = 1, 2, \ldots, L$$
$$\eta_r^{Room} \sim N\left(0, \left(\sigma_\eta^{Room}\right)^2\right), \text{ for } r = 1, 2, \ldots, R$$

where

- y_i is the ith observed `log(price)` in a location (or neighborhood) $l[i]$ with unit of $r[i]$ room type;
- L is the total number of locations in Toronto, here represented by `neighborhood`, and R is the number of `room_type`'s;
- $x_{i,1}$ is number of `bedrooms` of ith observation;
- $x_{i,2}$ is number of `accommodates` of ith observation;
- $x_{i,3}$ is number of `bathrooms` of ith observation;
- $\eta_{l[i]}^{Location}$ is the coefficient of influence from the lth location of the ith unit;
- $\eta_{r[i]}^{Room}$ is the coefficient of influence from the rth room type of the ith unit

IMPORTANT NOTE: Due to computational limitations, I had to restrict the database to a sample of not more than 20% from the original database. The `rstan`-model with $1,500$ iterations and the whole database consumed something around 5-6 hours to process, which is incompatible with the compressed time frame we have to develop and test the solutions present on the exam. Despite of that, I can assume with a certain degree of confidence that the estimates and adherence of the model are not harmed by this limitation.

[3]In fact I tested 04 models to reach tho these 02-finalists, one producing a design matrix to include `neighborhood` and `room_type` as categorical variables and producing a lot of *dummies* as we have 140 neighborhoods present in the database and 04 types of unit. I rather preferred concentrate on include just such complexity on the second model, leaving the first as simple as possible. So, that's my Model 2 is identified as "Model 4" in `.rmd` file.

Model 1

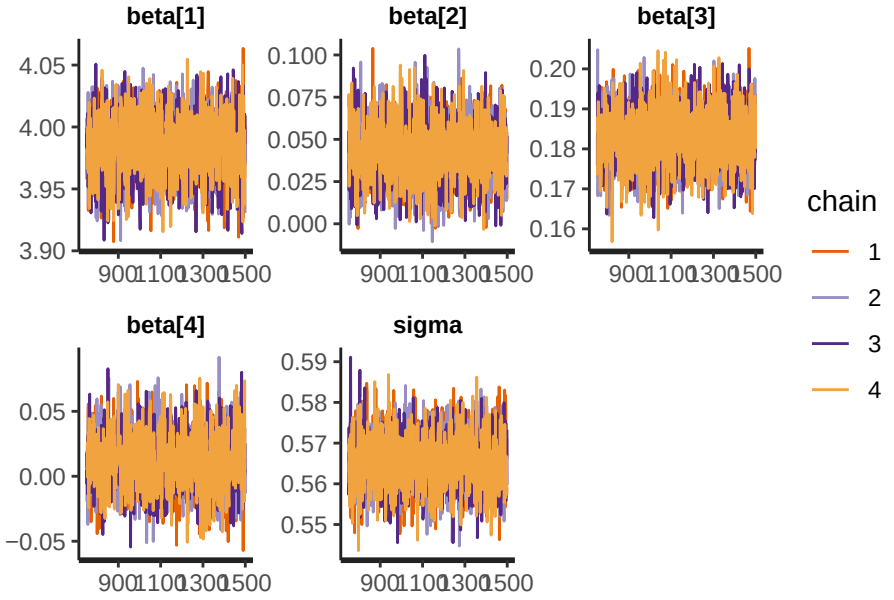


Figure 17: Diagnostics from Model 1 - Traceplots

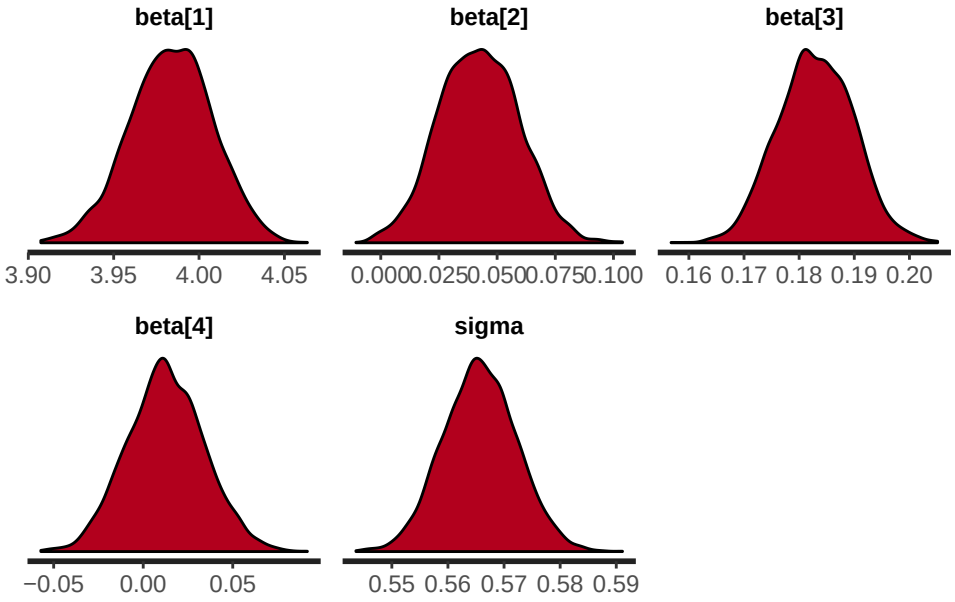


Figure 18: Diagnostics from Model 1 - Parameter Densities

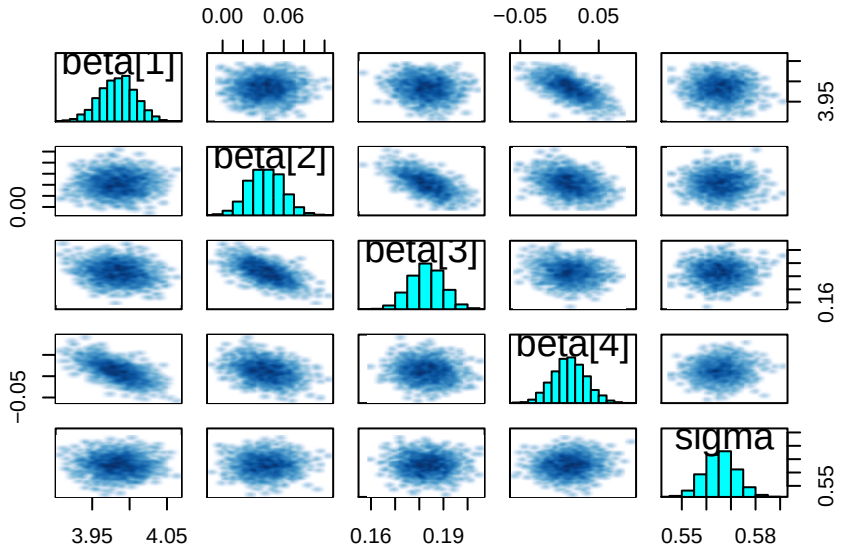


Figure 19: Diagnostics from Model 1 - Parameters Pair-Plots

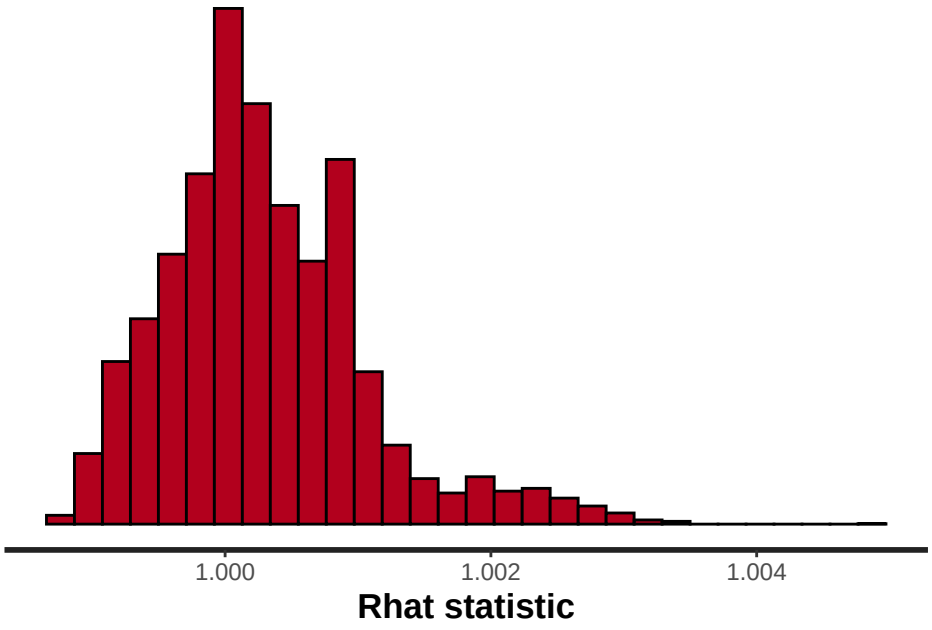


Figure 20: Diagnostics from Model 1 - RHat

The diagnostics from **Model 1** shows there is a good fit, the model converge, chains are well mixed and almost all RHat's are distributed around 1.0.

Model 2

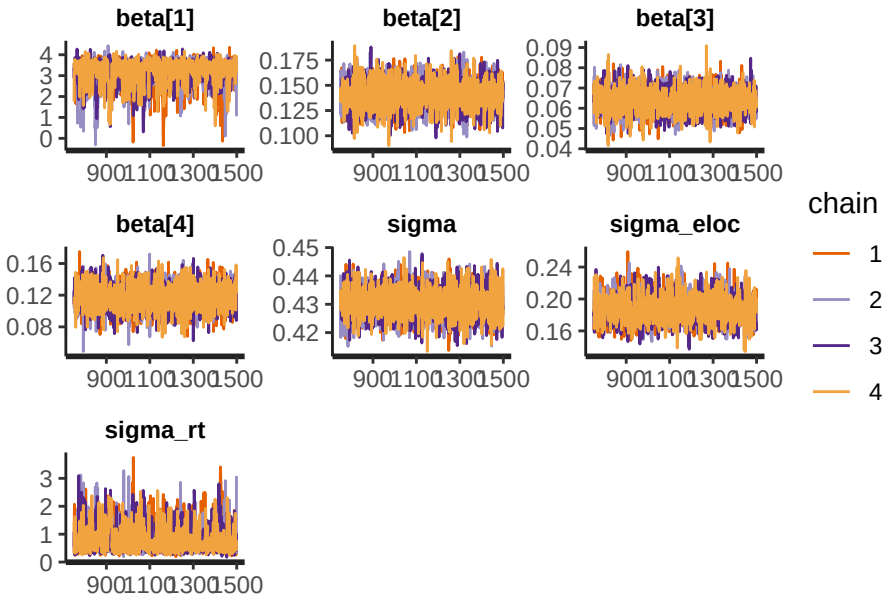


Figure 21: Diagnostics from Model 2 - Traceplots

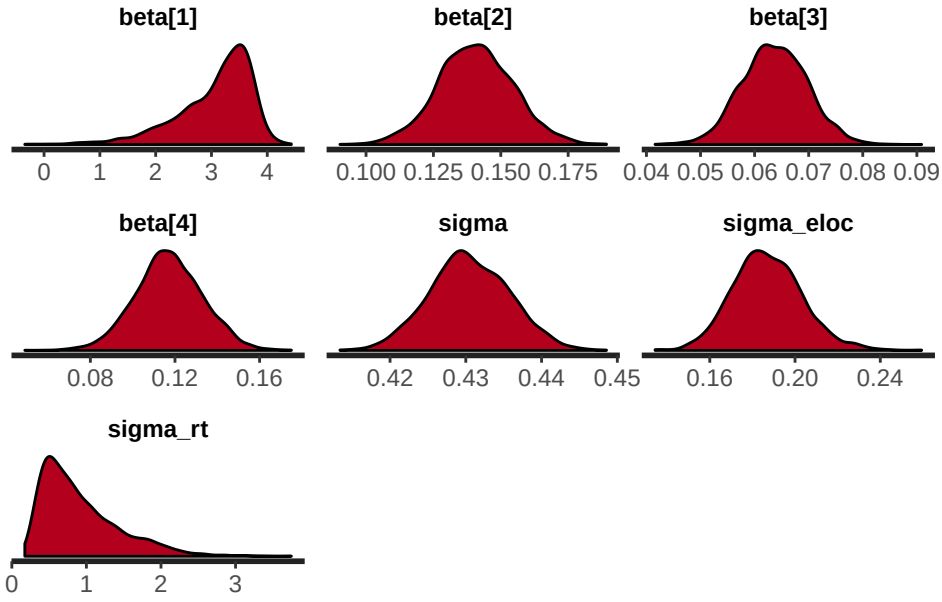


Figure 22: Diagnostics from Model 2 - Parameters Densities

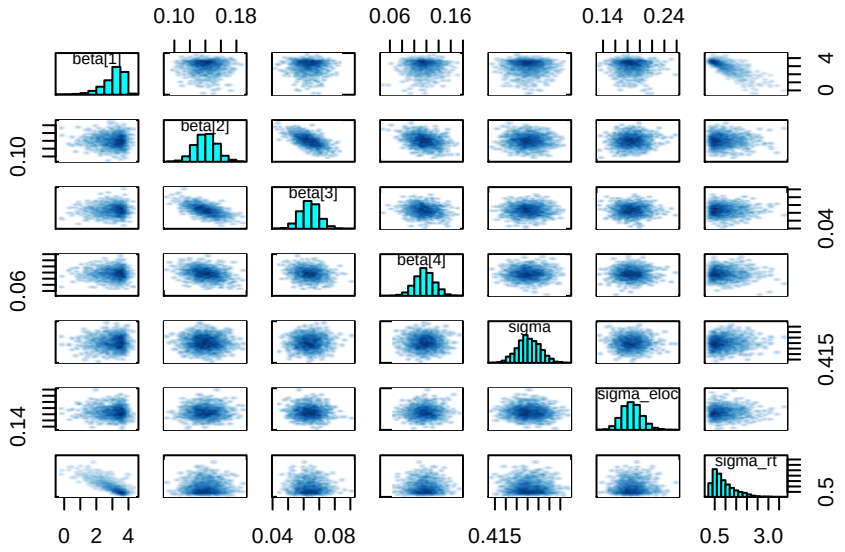


Figure 23: Diagnostics from Model 2 - Parameters Pair-plots

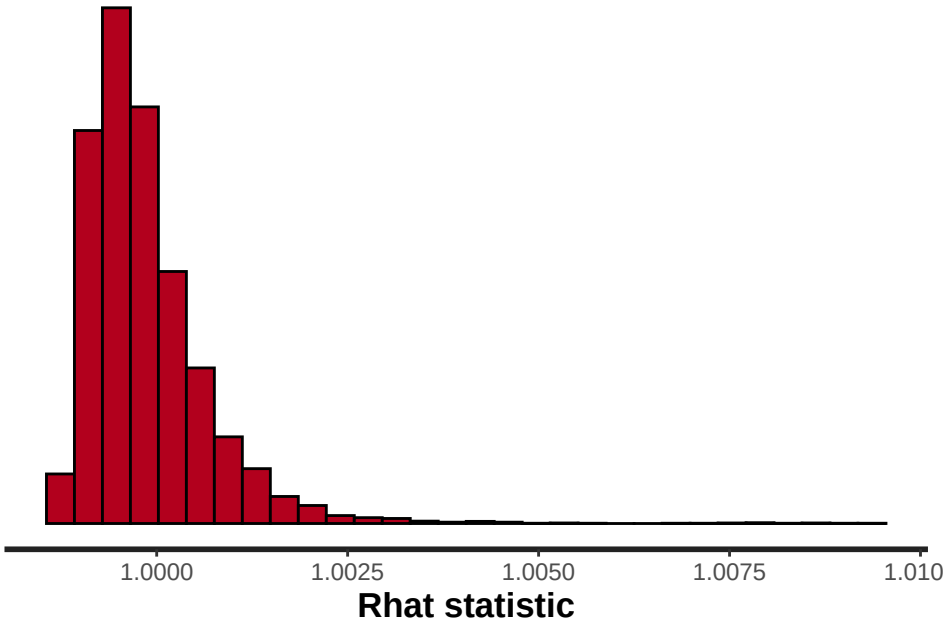


Figure 24: Diagnostics from Model 2 - RHat

The diagnostics of Model 2 has proven it is a fair fit, with **RHat** distributed around 1.0, the chains mixed well, but some pairs of parameters may present a slight correlation. Despite of this, the model was not rejected by the thresholds set to STAN so we will proceed with this model to compare with Model 1.

(c) Using LOO-CV, determine which of your models is preferred.

{*Answer.*}

By simulating LOO process to calculate the ELPD for both models, we have the following results:

```
##
## Computed from 3000 by 3696 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo  -3142.6  65.8
## p_loo         9.7   1.4
## looic      6285.2 131.7
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
##
## Computed from 3000 by 3696 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo  -2181.2  84.4
## p_loo        98.7   5.4
## looic      4362.5 168.9
## ------
## Monte Carlo SE of elpd_loo is 0.2.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     3694  99.9%   400
##  (0.5, 0.7]   (ok)          2   0.1%   520
##    (0.7, 1]   (bad)         0   0.0%   <NA>
##    (1, Inf)   (very bad)    0   0.0%   <NA>
##
## All Pareto k estimates are ok (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

By checking the estimated density and by selecting a sample of 20% of simulations generated and plot it against the observed distribution of y_i.

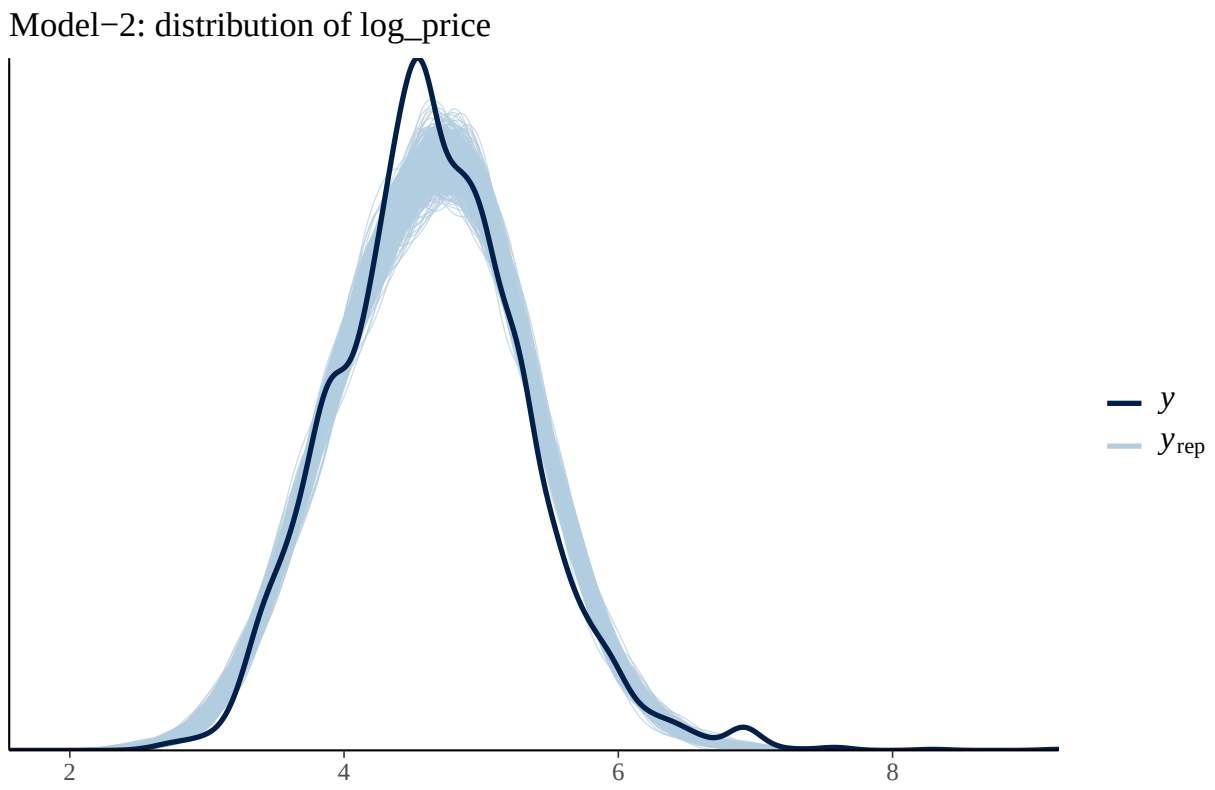


Figure 25: Comparison - Densities of Simulated vs. Observed data

We can see that simulated `y_rep` has a good fit for the model chosen.

In our case, **Model 2** has the lowest LOO-IC, which means the highest $elpd_Loo$, so we can consider it as the preferred model.

```
##        elpd_diff se_diff
## model2    0.0       0.0
## model1 -961.4      45.8
```

The `loo_compare` procedure has then confirmed **Model 2** is the best model.

(d) For the preferred model, discuss the results with the help of good explanations and graphs. It's up to you what you highlight, but just note that discussing values of coefficients from `summary(mod)` is not enough.

{*Answer.*}

Model 2 have some particularities when compared with **Model 1** as it considers 02 additional aspects not present in *Model 1*: 1) the differences of rates due to type of unit; and 2) which region in Toronto this units is located. It is evident that those 02 information are relevant and have influence over the nightly rate.

Let's so take a look at the coefficients and look for other interesting characteristics of the model.

Table 3: Coefficients for Model 2

	mean	se_mean	sd	Rhat
beta[1]	3.0635703	0.0326250	0.6771639	1.0093459
beta[2]	0.1406554	0.0002209	0.0138590	0.9990034
beta[3]	0.0637153	0.0000874	0.0060476	0.9989161
beta[4]	0.1171202	0.0002305	0.0159678	0.9994124

Table 4: Variances for Model 2

	mean	se_mean	sd	Rhat
sigma	0.4304979	0.0000699	0.0052478	0.9994603
sigma_eloc	0.1870518	0.0003658	0.0163359	1.0021714
sigma_rt	0.9139068	0.0222301	0.5180847	1.0074425

We can summarize the preliminary findings as follows:

- *Constant term*: the price for rental in Toronto starts from a level represented by the independent coefficient, i.e., in this case $e^{3.0635} = \$21.4$;
- *Coefficient for Unit Characteristics*: The coefficients for `Bathrooms` besides `Bedrooms` are the most influential characteristic of the unit. These coefficients may vary depending of room type, for example, for hotel rooms, number of bathrooms is not a good predictor because it is implicit every room having its own bathroom. For shared rooms, number of accommodates might be have more influence and so on.

Table 5: Coefficients for Room Type

	room_type	mean	se_mean	sd	Rhat
eta_rt[1]	Entire home/apt	1.0542671	0.0325881	0.6769493	1.008759
eta_rt[2]	Hotel room	0.9883975	0.0324883	0.6844604	1.008218
eta_rt[3]	Private room	0.4861788	0.0325495	0.6766212	1.008923
eta_rt[4]	Shared room	0.1000651	0.0323856	0.6754132	1.008608

For `room_type` it is evident that the price has more influence on type *Entire Home/Apt* because it is implicit that this type of unit is more expensive when compared with other units. On the other extreme we have *Shared Room* which may include university residences with more than one room mate, hostels etc. Observing the values of the coefficients, we may conclude that *Hotel Rooms* are approximately -6.4% cheaper than *Entire Home/Apt*. In the same sense, *Private Rooms* are -39.5% cheaper than *Hotel Rooms* and finally, *Shared Rooms* are usually -32.0% cheaper than *Private Rooms*.

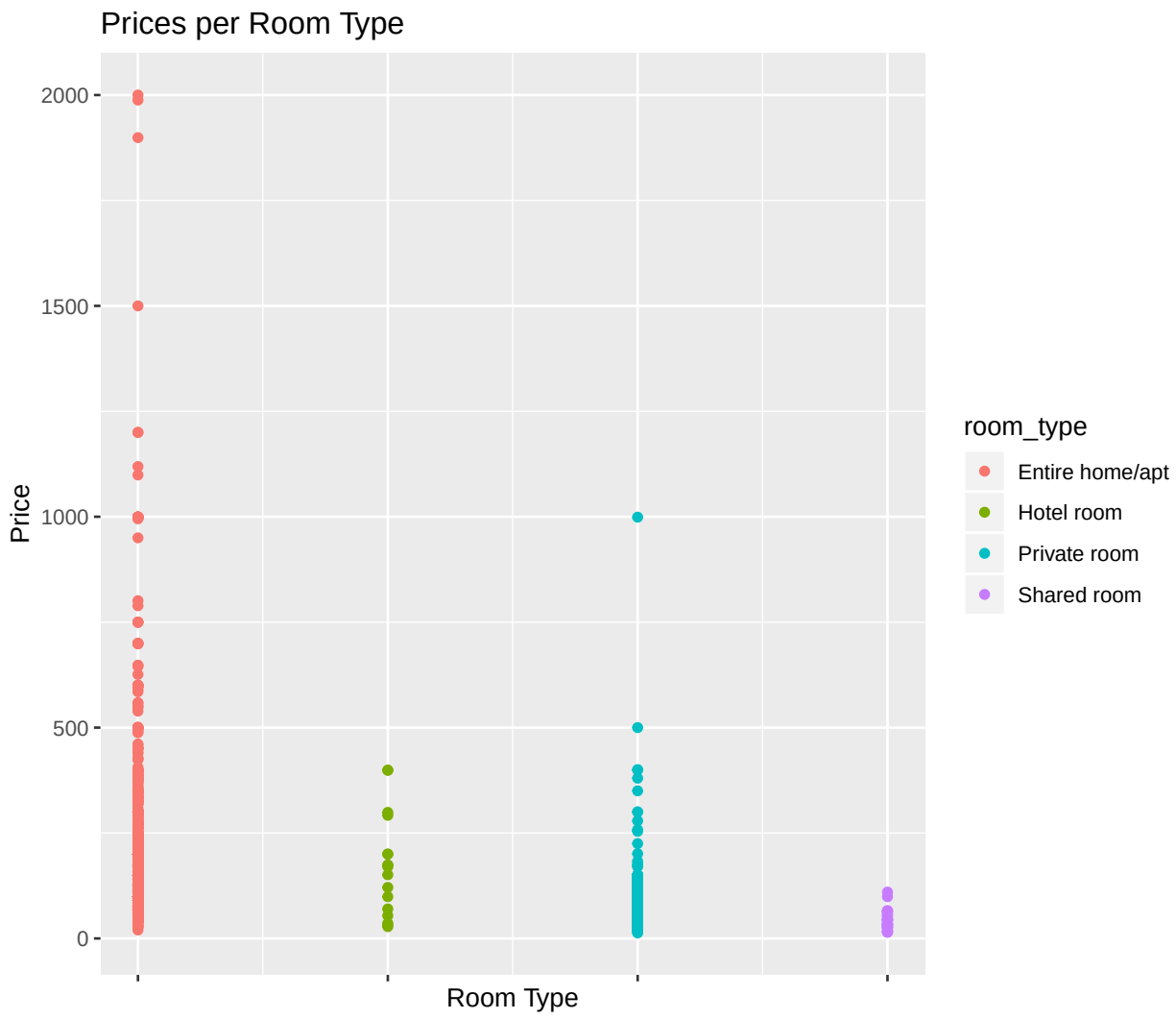


Figure 26: Prices per Room Type

Table 6: Coefficients for units with increasing daily-rate due to neighbourhood

neighb	mean	se_mean	sd	Rhat	incpct
Waterfront Communities-The Island	0.4854665	0.0008476	0.0271937	1.0078580	62.49
Bay Street Corridor	0.4063596	0.0009588	0.0467982	1.0039995	50.13
Niagara	0.4022924	0.0009598	0.0384226	1.0058095	49.52
Yonge-St.Clair	0.3037369	0.0016494	0.1082456	1.0006783	35.49
Moss Park	0.2887444	0.0009460	0.0515945	1.0011773	33.48
Annex	0.2865809	0.0009331	0.0411726	1.0043124	33.19
Church-Yonge Corridor	0.2826713	0.0009155	0.0439989	1.0024650	32.67
Cabbagetown-South St.James Town	0.2635144	0.0010514	0.0626596	1.0015552	30.15
Yonge-Eglinton	0.2544741	0.0012261	0.0869629	1.0009489	28.98
Regent Park	0.2433021	0.0015307	0.1024460	0.9999690	27.55
Mount Pleasant West	0.2416140	0.0011306	0.0685711	1.0006126	27.33
Rosedale-Moore Park	0.2405527	0.0014318	0.0923582	1.0003683	27.20
Trinity-Bellwoods	0.2399781	0.0010465	0.0501837	1.0023053	27.12
Kensington-Chinatown	0.2207993	0.0009221	0.0487747	1.0008467	24.71
Bayview Village	0.2140664	0.0014943	0.1000194	1.0000424	23.87
Casa Loma	0.2025335	0.0012701	0.0900004	1.0005961	22.45
Palmerston-Little Italy	0.2006618	0.0010165	0.0516538	1.0029724	22.22
North St.James Town	0.1990696	0.0011882	0.0683236	1.0008745	22.03
Little Portugal	0.1988934	0.0010320	0.0478059	1.0031940	22.01
Oakridge	0.1957677	0.0017364	0.1168171	0.9993710	21.62
North Riverdale	0.1793775	0.0012244	0.0758250	0.9994773	19.65
University	0.1719666	0.0011852	0.0763646	1.0003361	18.76
Mount Pleasant East	0.1499616	0.0016031	0.1211546	0.9989050	16.18
Edenbridge-Humber Valley	0.1487443	0.0021416	0.1550603	0.9991287	16.04
High Park-Swansea	0.1313232	0.0011069	0.0673389	1.0003231	14.03
New Toronto	0.1205763	0.0014317	0.0990720	1.0006489	12.81
South Parkdale	0.1152578	0.0009969	0.0552910	1.0016170	12.22
South Riverdale	0.1071719	0.0009861	0.0508182	1.0014488	11.31
Centennial Scarborough	0.1022837	0.0019137	0.1520640	0.9993566	10.77
Old East York	0.0988485	0.0015864	0.1246574	0.9988232	10.39
The Beaches	0.0982862	0.0011103	0.0699447	1.0001601	10.33
Dufferin Grove	0.0978952	0.0010418	0.0621608	1.0012622	10.28

Table 7: Coefficients for units with decreasing daily-rate due to neighbourhood

neighb	mean	se_mean	sd	Rhat	incpct
Malvern	-0.3393270	0.0013720	0.0966667	0.9992277	-28.78
Keelesdale-Eglinton West	-0.2827508	0.0015815	0.1152297	0.9997514	-24.63
Tam O'Shanter-Sullivan	-0.2793619	0.0011272	0.0923163	0.9992383	-24.37
Black Creek	-0.2628753	0.0021151	0.1607785	0.9993661	-23.12
Agincourt South-Malvern West	-0.2605133	0.0013807	0.1027488	0.9992028	-22.93
Agincourt North	-0.2555862	0.0015878	0.1149703	1.0002477	-22.55
Milliken	-0.2217218	0.0014481	0.1126718	0.9992726	-19.89
Highland Creek	-0.1938206	0.0017587	0.1406111	0.9990271	-17.62
Glenfield-Jane Heights	-0.1875175	0.0015054	0.1152567	0.9992771	-17.10
Hillcrest Village	-0.1748785	0.0013676	0.0994681	0.9995325	-16.04
L'Amoreaux	-0.1706409	0.0012978	0.0971058	0.9994600	-15.69
Thistletown-Beaumond Heights	-0.1705742	0.0021592	0.1612681	0.9992947	-15.68
Weston	-0.1509618	0.0015888	0.1196110	0.9990486	-14.01
Downsview-Roding-CFB	-0.1481657	0.0014301	0.0945590	1.0012191	-13.77
Taylor-Massey	-0.1473881	0.0015139	0.1235858	0.9992210	-13.70
Willowdale West	-0.1467839	0.0011258	0.0805510	1.0000309	-13.65
Bedford Park-Nortown	-0.1466517	0.0016532	0.1291177	1.0000720	-13.64
Don Valley Village	-0.1452518	0.0012054	0.0858684	0.9990977	-13.52
Bendale	-0.1435409	0.0014155	0.1078069	1.0004970	-13.37
Cliffcrest	-0.1387636	0.0014480	0.1086673	0.9988800	-12.96
Ionview	-0.1366367	0.0016453	0.1224783	0.9997861	-12.77
Rouge	-0.1358026	0.0014350	0.1189573	0.9990079	-12.70
Humber Heights-Westmount	-0.1347435	0.0017873	0.1375295	0.9996443	-12.61
York University Heights	-0.1346945	0.0010866	0.0838025	0.9999992	-12.60
Oakwood Village	-0.1343760	0.0011707	0.0818358	0.9995211	-12.57
Kennedy Park	-0.1329934	0.0017470	0.1313437	0.9995338	-12.45
Kingsview Village-The Westway	-0.1294385	0.0016365	0.1309140	0.9993064	-12.14
Rockcliffe-Smythe	-0.1286450	0.0014086	0.1145750	0.9991368	-12.07
Wexford/Maryvale	-0.1283270	0.0011851	0.0885220	0.9993146	-12.04
Broadview North	-0.1249847	0.0017010	0.1270147	0.9991121	-11.75
O'Connor-Parkview	-0.1236620	0.0016053	0.1088779	0.9995534	-11.63
Eglinton East	-0.1148195	0.0017991	0.1460598	0.9989473	-10.85
Willowridge-Martingrove-Richview	-0.1142839	0.0013338	0.0969067	0.9993264	-10.80
Pleasant View	-0.1142479	0.0016230	0.1216470	0.9992957	-10.80
Caledonia-Fairbank	-0.1122877	0.0015378	0.1214174	0.9991614	-10.62
Brookhaven-Amesbury	-0.1121582	0.0015765	0.1208130	0.9990317	-10.61
Morningside	-0.1072604	0.0013888	0.1151588	0.9995799	-10.17
Birchcliffe-Cliffside	-0.1070025	0.0012618	0.0939119	0.9996346	-10.15

The location is the most diverse factor which influences the price of a nightly rental of a Unit in Toronto. The prices can be almost 63% just if one pick a neighborhood like *Waterfront Communities-The Island*, or 50.1% to stay at *Bay Street Corridor*. On the other hand, *Malvern*, *Keelesdale-Eglinton West* and *Tam O'Shanter-Sullivan* prices can be up to -28.8% cheaper as its contribution to price is much smaller than other neighborhoods.

Despite of this high rates, the majority of *Location*'s contributes to lower prices approximately -10.6%, in most cases.

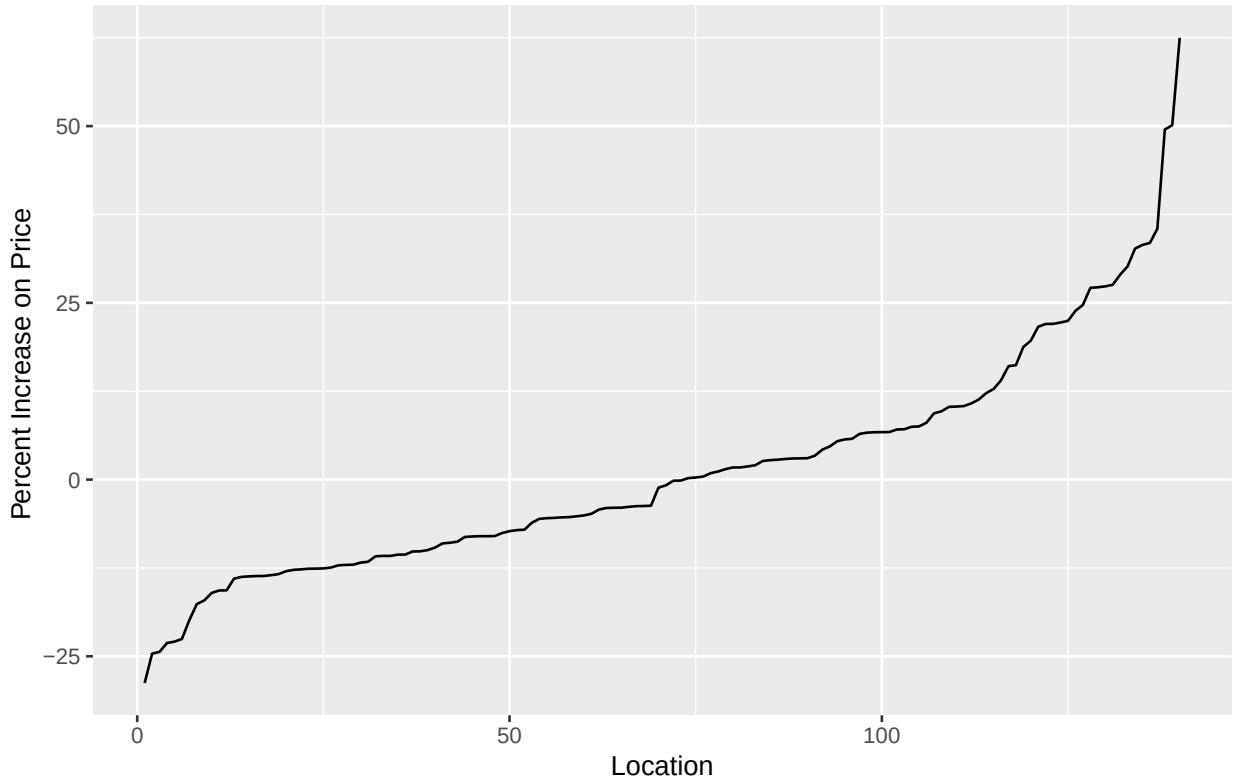


Figure 27: Contribution of Location to Rental Price

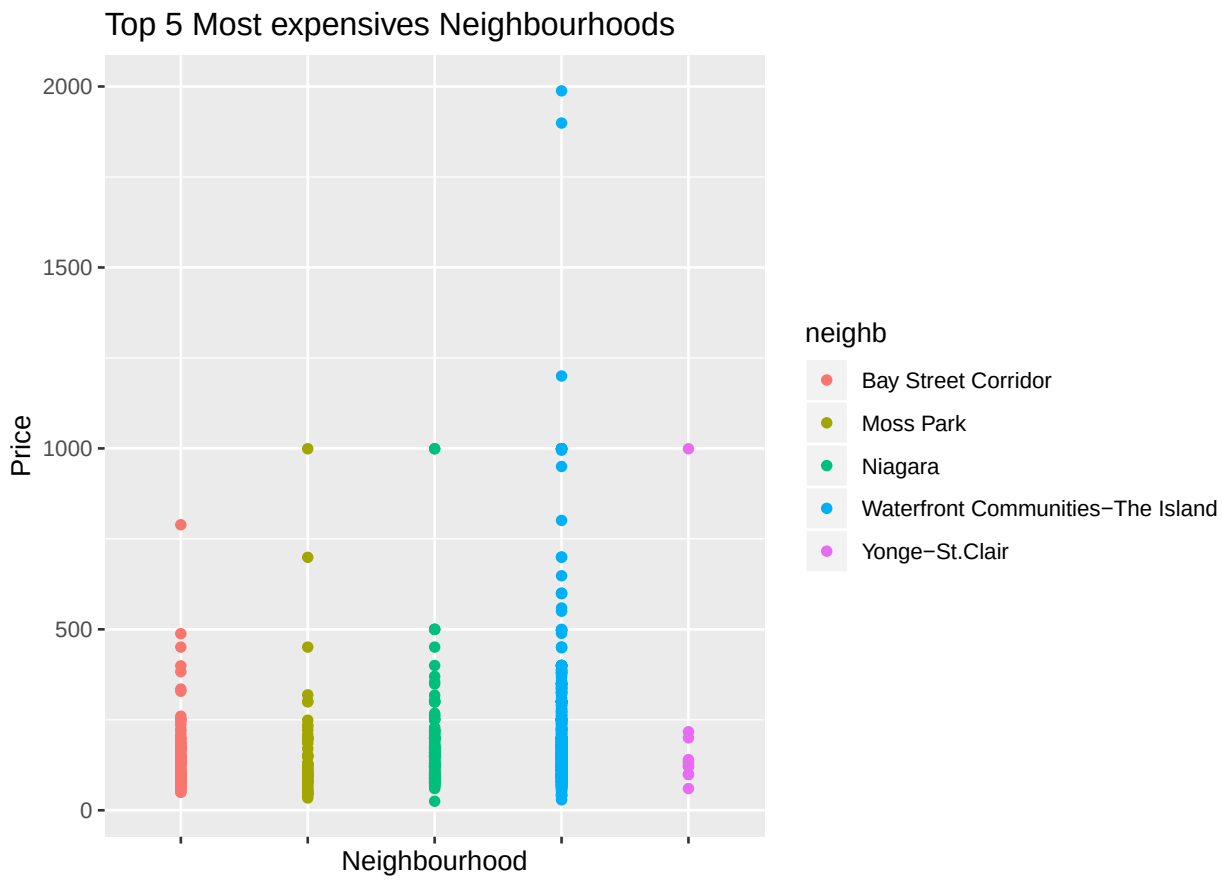


Figure 28: Highest Prices per Neighbourhood

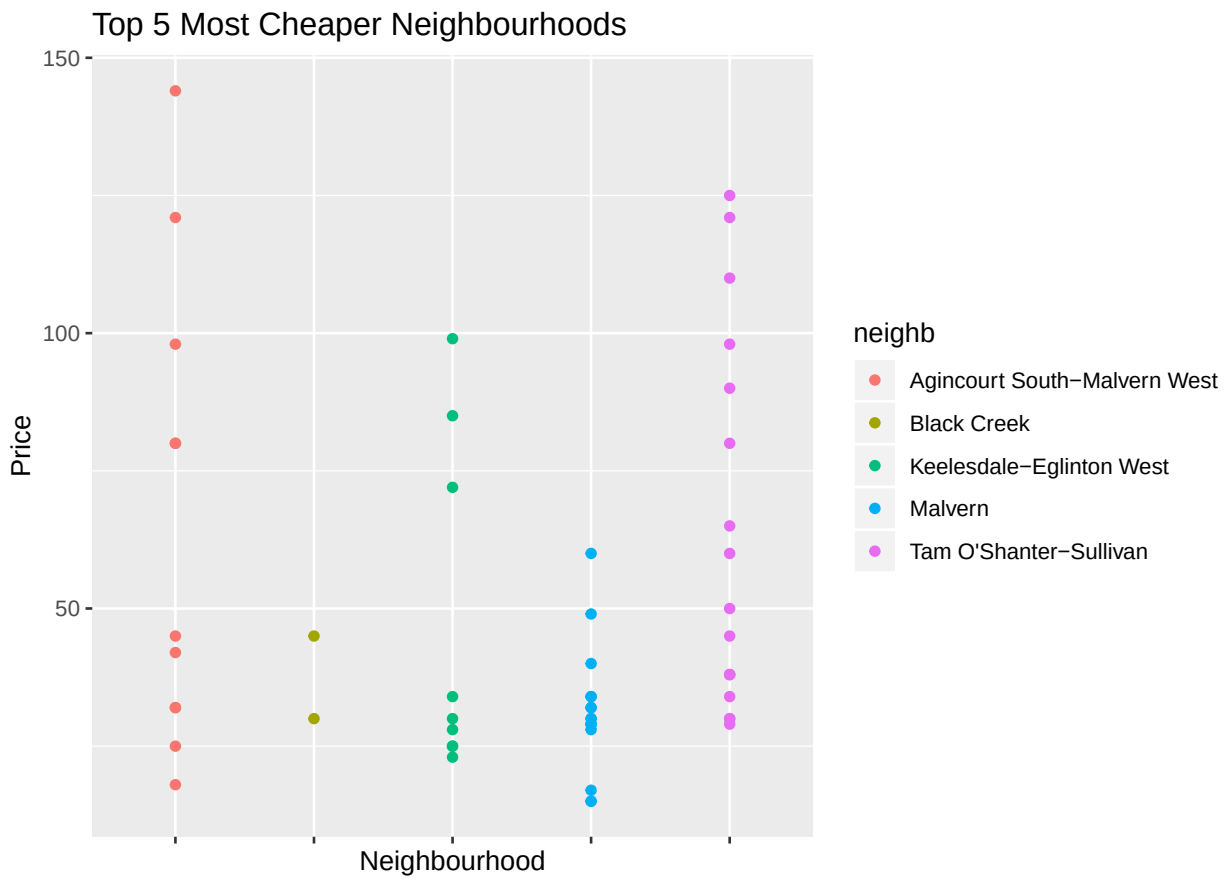


Figure 29: Lower Prices per Neighgourhood

I noticed the distribution of units is quite unbalanced between neighborhoods so this can influence the estimates $\eta^{location}$.

(e) Leave 20% of the data out at random (this is your test set). Rerun your preferred model on the remaining 80% of the data (this is the training set). Use the coefficient estimates to estimate the nightly rate for the each of the observations in the test set. The root mean squared error is:

$$RMSE = \sqrt{\frac{\sum_{i-1}^{N}(\hat{y_i} - y_i)^2}{N}}$$

where y_i is the observed `log(price)` and y_i is the estimated `log(price)`. Calculate the RMSE for the entire test set and also by room type. Briefly comment.

{*Answer.*}

Table 8: RMSE for Room Type - Test Database

room_type	RMSE
Entire home/apt	0.4423520
Hotel room	1.1164151
Private room	0.3871019
Shared room	0.7096357

Table 9: RMSE for Room Type - Training Database

room_type	RMSE
Entire home/apt	0.4332631
Hotel room	0.7944817
Private room	0.3984129
Shared room	0.3933444

Table 10: RMSE for Entire Database

RMSE
0.4246432

Table 11: Quantities of cases per room type - Test database

room_type	n
Entire home/apt	1238
Hotel room	4
Private room	587
Shared room	19

Table 12: Quantities of cases per room type - Training database

room_type	n
Entire home/apt	2481
Hotel room	17
Private room	1161
Shared room	37

COMMENTS:

We can observe that for those room types which have the greater presence on training set, the prediction is more accurate and, consequently the estimates are better. For rooms of type `Entire home/apt` the $RMSE_{Entirehome/apt}$ is near the $RMSE_{Tot}$ which is around 0.425. This behavior is observed probably because the frequency of this type of room is the largest in both databases (training and test). ON the other

hand, for type `Hotel`, the $RHMS_{Hotel}$ is larger as expected, as the frequency of this type of unit is the shortest one. The same behavior occurs on $RHMS_{Shared}$.

4 - Short Questions

(a) Show that if survival times are exponentially distributed, that the gamma distribution is the conjugate prior for the unknown hazard.

{*Answer.*}

Lets consider T be the survival time and let's assume it is distributed exponentially. Then we can then write:

$$T \sim Exp(\theta), \text{with } \theta > 0 \tag{8}$$

By (8) we can calculate likelihood function of θ by:

$$\begin{aligned} L(\theta; \mathbf{T}|\theta) &= \prod_{i=1}^{n} -\theta e^{-\theta t_i} \\ &= -\theta^n e^{-\theta \sum_{i=1}^{n} x_i} \\ &= P(\mathbf{T}|\theta) \end{aligned}$$

$$\implies P(\mathbf{T}|\theta) = -\theta^n e^{-\theta \sum_{i=1}^{n} x_i} \tag{9}$$

From (8) we also know that

$$\theta \sim Gamma(\alpha, \beta), \text{with } \alpha, \beta > 0$$

which implies

$$\implies P(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \tag{10}$$

We know from the basic *Rule of Bayes* that the posterior $P(\theta|\mathbf{T})$ is proportional to the prior and the likelihood, then:

$$P(\theta|\mathbf{X}) \propto L(\mathbf{T}|\theta) P(\theta) \tag{11}$$

From (9), (10) and (11), we have:

$$\begin{aligned} P(\theta|\mathbf{X}) &\propto L(\mathbf{T}|\theta) P(\theta) \\ &\propto -\theta^n e^{-\theta \sum_{i=1}^{n} x_i} \theta^{\alpha-1} e^{-\beta\theta} \\ &\propto -\theta^{\alpha+n-1} e^{-\theta(\sum_{i=1}^{n} x_i + \beta)} \end{aligned}$$

then, we conclude that

$$P(\theta|\mathbf{X}) \sim Gamma(\alpha + n, \sum_{i=1}^{n} x_i + \beta). \tag{12}$$

(b) Specify the likelihood function if you only observe $\bar{y} = 1/2(y_1 + y_2)$ where $y_i \sim N(\mu, \sigma^2)$ and $Cor(y_1, y_2) = \rho \neq 0$.

{*Answer.*}