Assignment 1

16340116

2024-07-02

Introduction

The aim of this assignment is to write a scientific report in which I use Bayesian workflow, Bayesian Hypothesis testing, Bayesian Model Selection, and Bayesian GLM to answer scientific questions related to a given dataset.

```
# Clear environment
rm(list=ls())

# Setting seed to ensure reproducible results
set.seed(1759)

# Loading required packages using pacman package
pacman::p_load(janitor,dplyr,tidyr,lubridate,stringr,ggplot2,readr,corrplot,
               bayesrules,rstanarm,bayesplot,tidyverse,tidybayes,broom.mixed,
               gridExtra)

# Reading in bikes data set
bikes <- read_csv("bikes2024.csv",col_types = cols())
```

Data

The dataset I will be analysing contains variables concerning bike traffic and weather conditions in Dublin in April and May 2024.

There are three variables concerning bicycle traffic volumes from cycle counters in four locations Dublin city: Clontarf, Griffith Avenue , Richmond Street and Grove Road. Passing cyclists are counted and logged every hour, 24 hours per day, 7 days per week. Data provided by Dublin City Council and the NTA in April and May 2024. The other variables concern weather condition, and they have been downloaded from Met Éireann.

The dataset consists of the following variables

1. *Date* and *Hour*: Timestamp for when observation data was collected.
2. *Day*: Day of the week.
3. *Clontarf* hourly bicycle traffic volumes from cycle counters in Clontarf (Pebble Beach Carpark).
4. *Griffith_Avenue* hourly bicycle traffic volumes from cycle counters in Griffith Avenue (Clare Rd Side).
5. *Richmond_Street* hourly bicycle traffic volumes from cycle counters in Richmond street (Inbound totem).
6. *Grove_Road* hourly bicycle traffic volumes from cycle counters in Grove Road Totem.
7. *rain* precipitation Amount (mm)

8. *temp* air Temperature (Degrees Celsius)
9. *wdsp* mean hourly wind speed (kt)
10. *vis* visibility (m)
11. *clamt* cloud amount (okta)

Initial Data Exlporation

```
# Initial data manipulation
bikes$Day <- factor(bikes$Day,
                    levels = c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"),
                    ordered = FALSE)

bikes$clamt <- factor(bikes$clamt, ordered = TRUE)

summary(bikes)
```

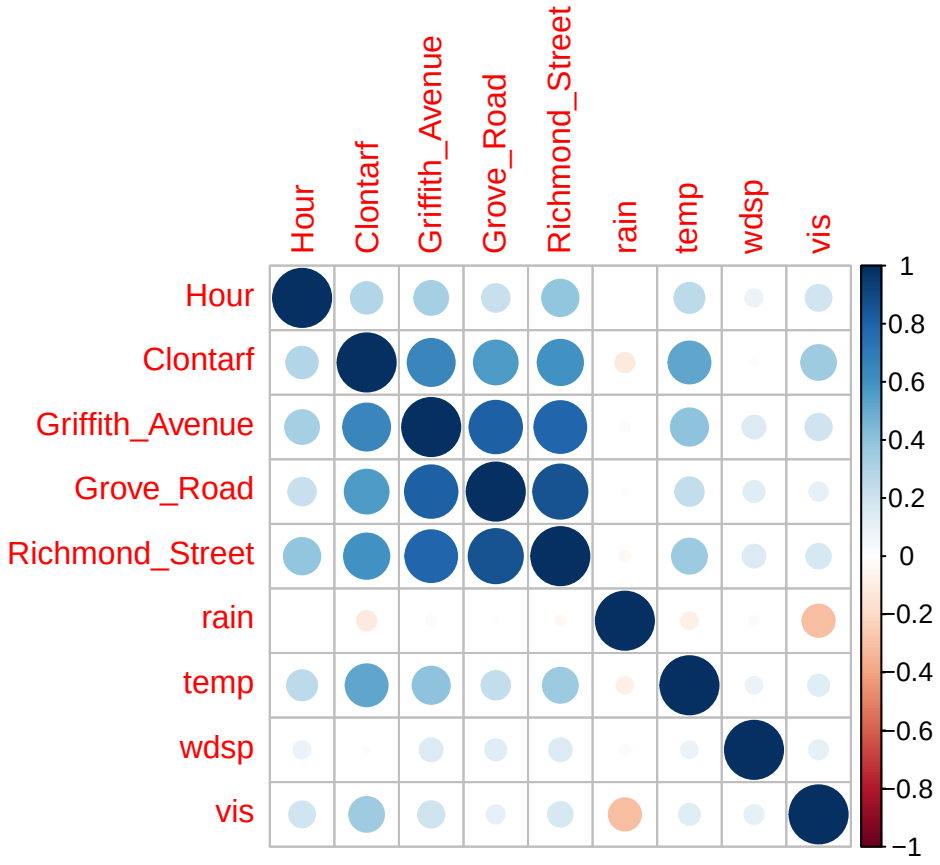
```
##       Date                 Hour        Day         Clontarf
##  Min.   :2024-04-01   Min.   : 0.00   Mon:216   Min.   :  0.00
##  1st Qu.:2024-04-16   1st Qu.: 5.75   Tue:216   1st Qu.:  8.00
##  Median :2024-05-01   Median :11.50   Wed:216   Median : 46.00
##  Mean   :2024-05-01   Mean   :11.50   Thu:216   Mean   : 59.82
##  3rd Qu.:2024-05-16   3rd Qu.:17.25   Fri:216   3rd Qu.: 86.00
##  Max.   :2024-05-31   Max.   :23.00   Sat:192   Max.   :365.00
##                                       Sun:192
##  Griffith_Avenue   Grove_Road    Richmond_Street       rain
##  Min.   : 0.0    Min.   :  0.0   Min.   :  0.00   Min.   :0.00000
##  1st Qu.: 2.0    1st Qu.: 30.0   1st Qu.: 13.00   1st Qu.:0.00000
##  Median :10.0    Median : 87.0   Median : 58.00   Median :0.00000
##  Mean   :13.1    Mean   :114.8   Mean   : 56.64   Mean   :0.07309
##  3rd Qu.:19.0    3rd Qu.:134.2   3rd Qu.: 83.00   3rd Qu.:0.00000
##  Max.   :69.0    Max.   :775.0   Max.   :283.00   Max.   :5.30000
##
##       temp            wdsp            vis            clamt
##  Min.   :-1.10   Min.   : 1.00   Min.   :  200   7      :642
##  1st Qu.: 8.60   1st Qu.: 6.00   1st Qu.:20000   6      :248
##  Median :11.10   Median : 8.00   Median :25000   8      :126
##  Mean   :10.89   Mean   : 9.02   Mean   :24320   5      :117
##  3rd Qu.:13.30   3rd Qu.:11.00   3rd Qu.:30000   1      :109
##  Max.   :21.20   Max.   :28.00   Max.   :55000   3      : 94
##                                                  (Other):128
```

Above, I display some high-level descriptive statistics for each of the variables in the 'bikes' dataset. I also produce a correlation-plot of all the numeric variables in the dataset below. We observe strong positive linear correlation between the 4 locations in Dublin namely (Clontarf, Griffith Avenue, Grove Road and Richmond Street) which suggests that when bike traffic volumes are high in one area, they tend to be high in the other areas and vice-versa.

We observe hardly any linear correlation between rain or windspeed and the bike traffic volumes in the city which I find surprising. We do observe weakly positive relationships between temperature and visibility with bike traffic in the 4 locations.

```
corr_bikes <- bikes %>%
  select(-Date,-Day,-clamt)

corrplot(cor(corr_bikes))
```



Question 1

Is the average temperature in April and May less than 11 degrees celsius? (Assuming known variance of $\sigma^2 = 9$*)*

To answer the above question, I will use Bayesian Hypothesis testing similar to the example provided in the lecture notes about Download Speeds. Firstly, I plot average daily temperatures. I also calculate the average temperature in April/May to be 10.89 degrees which is slightly less than 11 degrees.

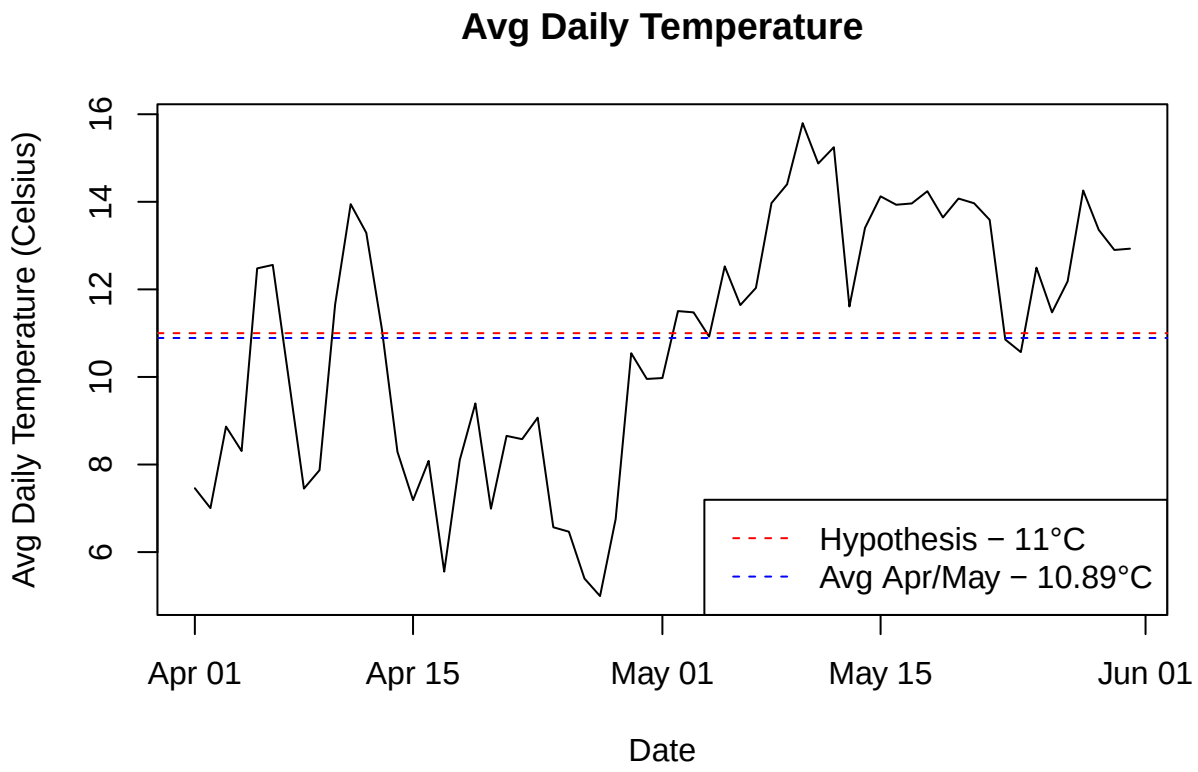
I use Normal-Normal conjugacy pairs for prior and likelihood models. As we are using time series data the iid assumption breaks down however, per the discussion board and for the purposes of this analysis, I assume that it doesn't violate the normal-normal conjugacy assumptions.

```
# plotting average daily temperatures for April and May
temp_plot <- bikes %>%
  select(Date,temp) %>%
  group_by(Date,) %>%
  mutate(avg_day_temp = mean(temp)) %>%
  ungroup %>%
  select(-temp) %>%
```

```r
  distinct()

# Avg daily temp April & May
avg_temp_apr_may <- round(mean(temp_plot$avg_day_temp),2)

plot(temp_plot$Date,temp_plot$avg_day_temp,type = "l",lty = 1,
    xlab = "Date",ylab = "Avg Daily Temperature (Celsius)",
    main = "Avg Daily Temperature")
abline(h = 11, col = "red", lty = 2)
abline(h = avg_temp_apr_may, col = "blue", lty = 2)
legend("bottomright",
       legend = c("Hypothesis - 11\u00B0C","Avg Apr/May - 10.89\u00B0C"),
       lty = c(2,2),col = c("red","blue"))
```



I now formalise the hypothesis in order to determine if the average temperature in April/May is less than 11 degrees celsius.

$H_0 : \theta >= 11$ vs $H_1 : \theta < 11$

Model: $x_1, ..., x_{1464} | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2 = 9)$

Priors: I have used two normal distributions for my prior models for each hypothesis.

$\theta | H_0 \sim N_{\theta \geq \theta_0}(\mu_0 = 11.5, \sigma^2 = 9)$

$\theta | H_1 \sim N_{\theta < \theta_0}(\mu_0 = 9.5, \sigma^2 = 9)$

For H0, I have used a mean of 11.5 which is greater than the hypothesised temperature used in the research question of 11 degrees celsius.

For H1, I have used a mean of 9.5 for my prior distribution of temperature based on the LTA (long term average) temperature in April & May in the Phoenix Park in Dublin provided by Met Eireann (Link).

The known variance of $\sigma = 9$ is used in both prior models, which I believe sufficiently reflects the uncertainty present in our prior beliefs.

```
# Calculating required components for Bayes' Factor Analysis
x <- bikes$temp
x_bar <- mean(bikes$temp)
n <- length(x)

theta_0 <- 11
sigma <- 9
mu_0 <- 11.5
mu_1 <- 9.5

# Calculating p(data,H0)
p_data_H0 <- function(theta) {
  # p(data|theta)
  dnorm(x_bar, mean = theta, sd = sqrt(sigma / n)) *
  # p(theta|H0)
  dnorm(theta, mean = mu_0, sd = sqrt(sigma))
}

# Calculating p(data,H1)
p_data_H1 <- function(theta) {
  # p(data|theta)
  dnorm(x_bar, mean = theta, sd = sqrt(sigma / n)) *
  # p(theta|H0)
  dnorm(theta, mean = mu_1, sd = sqrt(sigma))
}

# Calculating p(data|H0)
p_data_given_H0 <- integrate(p_data_H0,theta_0,Inf)$value

# Calculating p(data|H1)
p_data_given_H1 <- integrate(p_data_H1,-Inf,theta_0)$value

# Calculating Bayes' Factor
BF10 <- p_data_given_H1 / p_data_given_H0
BF10
```

```
## [1] 9.47872
```

Above I have calculated the Bayes' factor and we can see there is positive evidence to suggest that the average temperature in April / May is less than 11 degrees celsius ($\bar{x} = 10.89$).

Question 2

Is there a correlation or association between the number of cyclists passing on different locations on the same date?

As we are dealing with counts of cyclists, I will propose to use either a Poisson or Negative-Binomial GLM data model to answer this question. Firstly, I will aggregate the cyclist count data to daily values in order to compare the total number of cyclists passing each location on each day.

```r
# Aggregating cyclist counts to daily figures
bikes_q2 <- bikes %>%
  select(Clontarf,Griffith_Avenue,Richmond_Street,Grove_Road,Date) %>%
  group_by(Date) %>%
  mutate(Clontarf_Daily = sum(Clontarf),
         Griffith_Avenue_Daily = sum(Griffith_Avenue),
         Richmond_Street_Daily = sum(Richmond_Street),
         Grove_Road_Daily = sum(Grove_Road)) %>%
  ungroup() %>%
  select(-Clontarf,-Griffith_Avenue,-Richmond_Street,-Grove_Road) %>%
  unique()
```

I observe the distributions of each of the count data across the 4 locations in Dublin to help determine over-dispersion and assess normality of variables.

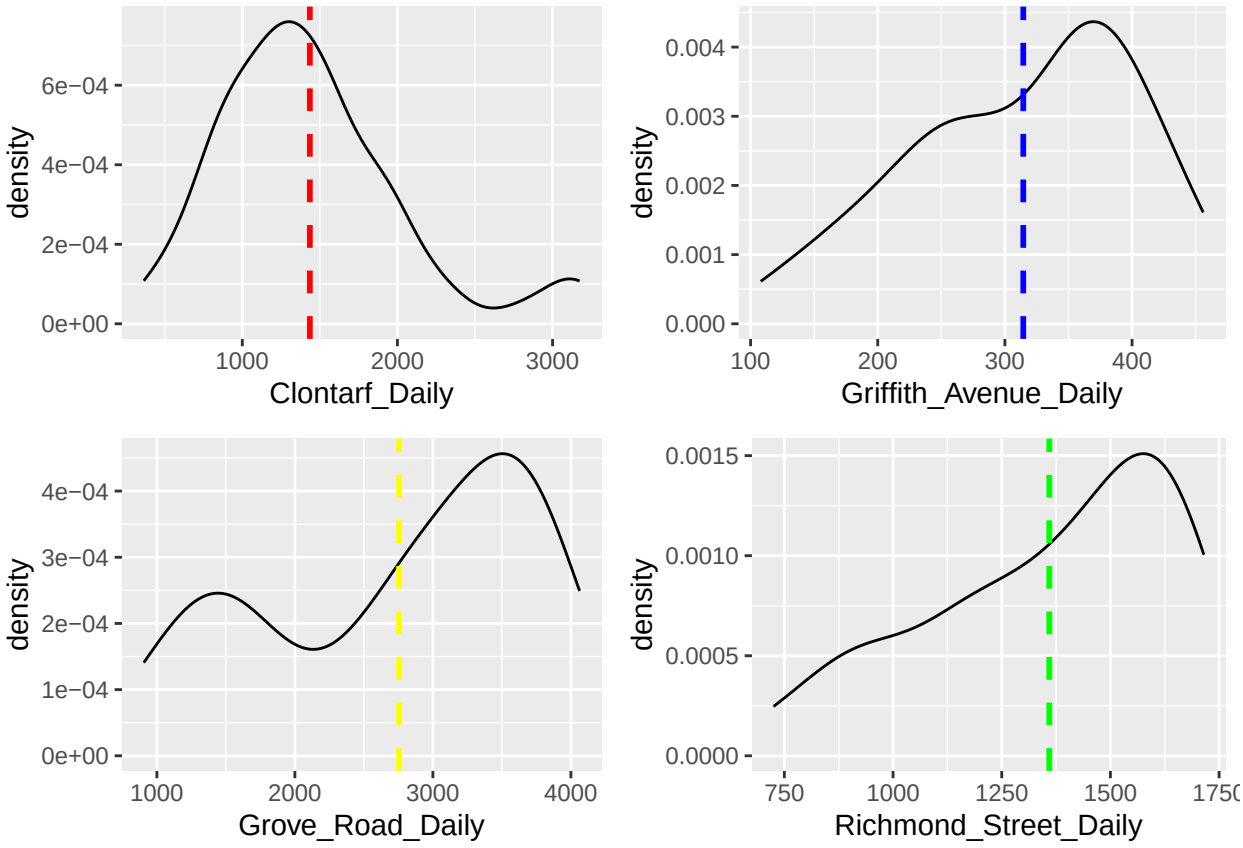
```r
# Checking distributions of target and response variables
p1 <- ggplot(bikes_q2,aes(x=Clontarf_Daily)) +
        geom_density() +
        geom_vline(aes(xintercept=mean(Clontarf_Daily)),
            color="red", linetype="dashed", linewidth=1)

p2 <- ggplot(bikes_q2,aes(x=Griffith_Avenue_Daily)) +
        geom_density() +
        geom_vline(aes(xintercept=mean(Griffith_Avenue_Daily)),
            color="blue", linetype="dashed", linewidth=1)

p3 <- ggplot(bikes_q2,aes(x=Grove_Road_Daily)) +
        geom_density() +
        geom_vline(aes(xintercept=mean(Grove_Road_Daily)),
            color="yellow", linetype="dashed", linewidth=1)

p4 <- ggplot(bikes_q2,aes(x=Richmond_Street_Daily)) +
        geom_density() +
        geom_vline(aes(xintercept=mean(Richmond_Street_Daily)),
            color="green", linetype="dashed", linewidth=1)

grid.arrange(p1,p2,p3,p4,nrow = 2, ncol = 2)
```



I have plotted each of the distributions of the number of cyclists passing each of the 4 Dublin locations daily above. As we can see the distribution of each of the locations is non-normal, over-dispersed and multi-modal in some cases. I will use a Negative-Binomial data model for this question in order to address the issue of over-dispersion.

I will also choose Grove Road as my target variable and the remaining 3 locations as explanatory variables in order to explore whether there is a relationship between the number of cyclists passing daily in different locations.

As the response variable is non-normal, we will not have conjugate priors for the beta coefficients and as such, I will use sufficiently conservative normal priors and assume independence between them. I do not believe that these coefficients will be independent based on the high correlation between these variables per my correlation plot previously however, for the purpose of this question I will assume independence.

As I am using a Negative-Binomial data model, we have that $\mu = exp(X\beta)$ and as such I will choose a prior for the intercept which reflects $log(\mu) = X\beta$. The mean value for Grove Road is 2755. Therefore I will use a $N \sim (8, 2)$ prior for the intercept as I believe this is a reasonable prior distribution.

```
# Estimating posterior using rstanarm
bikes_model_negbin_prior <- stan_glm(Grove_Road_Daily ~ Clontarf_Daily +
                                     Griffith_Avenue_Daily + Richmond_Street_Daily,
                                     data = bikes_q2,
                                     family = neg_binomial_2,
                                     #prior for beta_0
                                     prior_intercept = normal(8,2),
                                     #tune remaining priors
                                     prior = normal(0,4,autoscale = TRUE),
                                     #tune prior for r
```

```
                                    prior_aux = exponential(1, autoscale = TRUE),
                                    chains = 4,
                                    iter = 5000*2,
                                    seed = 1759,
                                    refresh = 0,
                                    prior_PD = TRUE)
```

```
# Estimate posterior using rstanarm
bikes_model_negbin_prior$prior.info
```

```
## $prior
## $prior$dist
## [1] "normal"
##
## $prior$location
## [1] 0 0 0
##
## $prior$scale
## [1] 4 4 4
##
## $prior$adjusted_scale
## [1] 0.006535369 0.044872632 0.014302878
##
## $prior$df
## NULL
##
##
## $prior_intercept
## $prior_intercept$dist
## [1] "normal"
##
## $prior_intercept$location
## [1] 8
##
## $prior_intercept$scale
## [1] 2
##
## $prior_intercept$adjusted_scale
## NULL
##
## $prior_intercept$df
## NULL
##
##
## $prior_aux
## $prior_aux$dist
## [1] "exponential"
##
## $prior_aux$location
## NULL
##
## $prior_aux$scale
## NULL
```

```
##
## $prior_aux$adjusted_scale
## NULL
##
## $prior_aux$df
## NULL
##
## $prior_aux$rate
## [1] 1
##
## $prior_aux$aux_name
## [1] "reciprocal_dispersion"
```

Based on the above output stan_glm gives the following priors:

$\beta_1 \sim Normal(0, 0.048^2)$, $\beta_2 \sim Normal(0, 0.232^2)$, $\beta_3 \sim Normal(0, 0.064^2)$,
$r \sim Exp(1)$

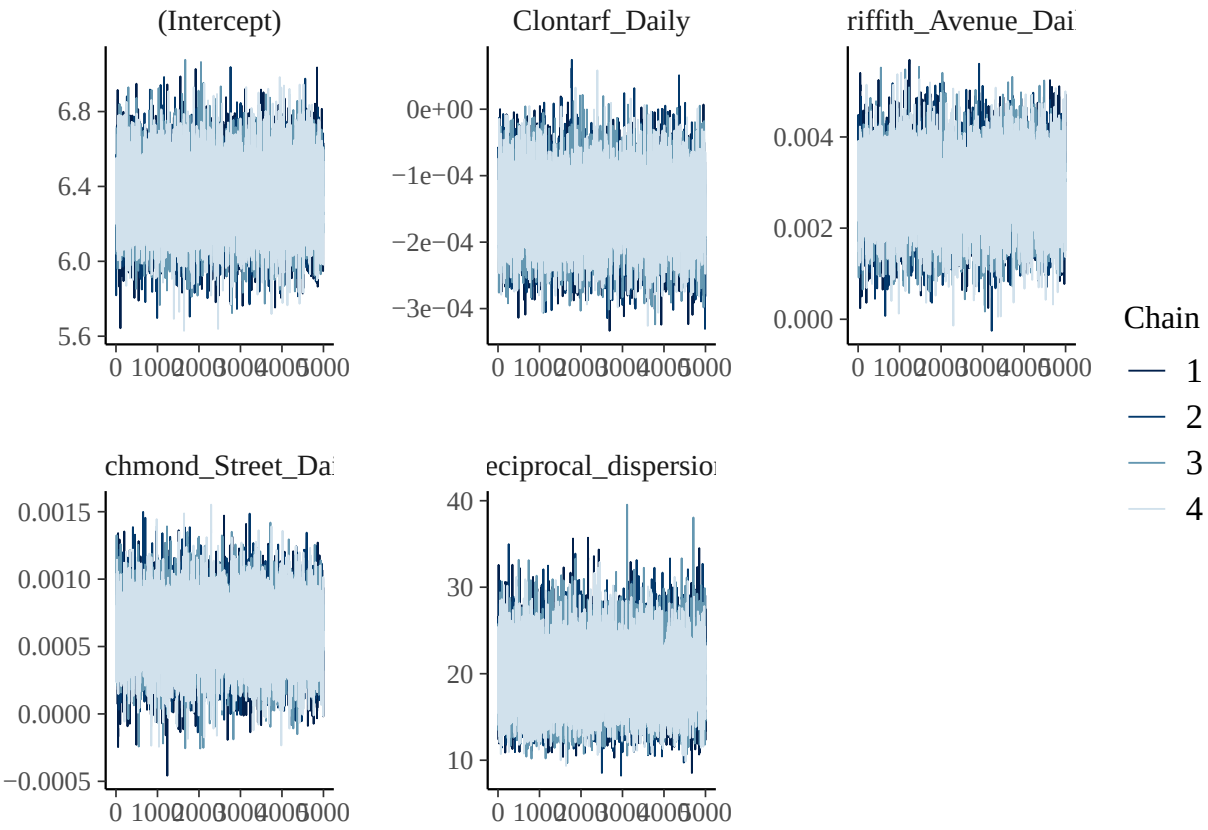
```
# Simulating the posterior
bikes_model_negbin <- update(bikes_model_negbin_prior, prior_PD = FALSE)
tidy(bikes_model_negbin, conf.int = TRUE, conf.level = 0.95)
```

```
## # A tibble: 4 x 5
##   term                   estimate std.error  conf.low  conf.high
##   <chr>                     <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)            6.34     0.175      6.00      6.70
## 2 Clontarf_Daily        -0.000148 0.0000492 -0.000246 -0.0000466
## 3 Griffith_Avenue_Daily  0.00280  0.000727   0.00135   0.00429
## 4 Richmond_Street_Daily  0.000616 0.000226   0.000152  0.00107
```

Above we observe the Bayesian regression output for the model. We can see there is a large intercept coefficient of 6.34 and is significant given that the 95% interval is far from (and does not contain) zero. Interestingly we observe a very small negative coefficient with the volume of Clontarf cyclists, a one unit increase in the Clontarf variable will result in a exp(0.000148) decrease in expected cyclists on Grove Road.

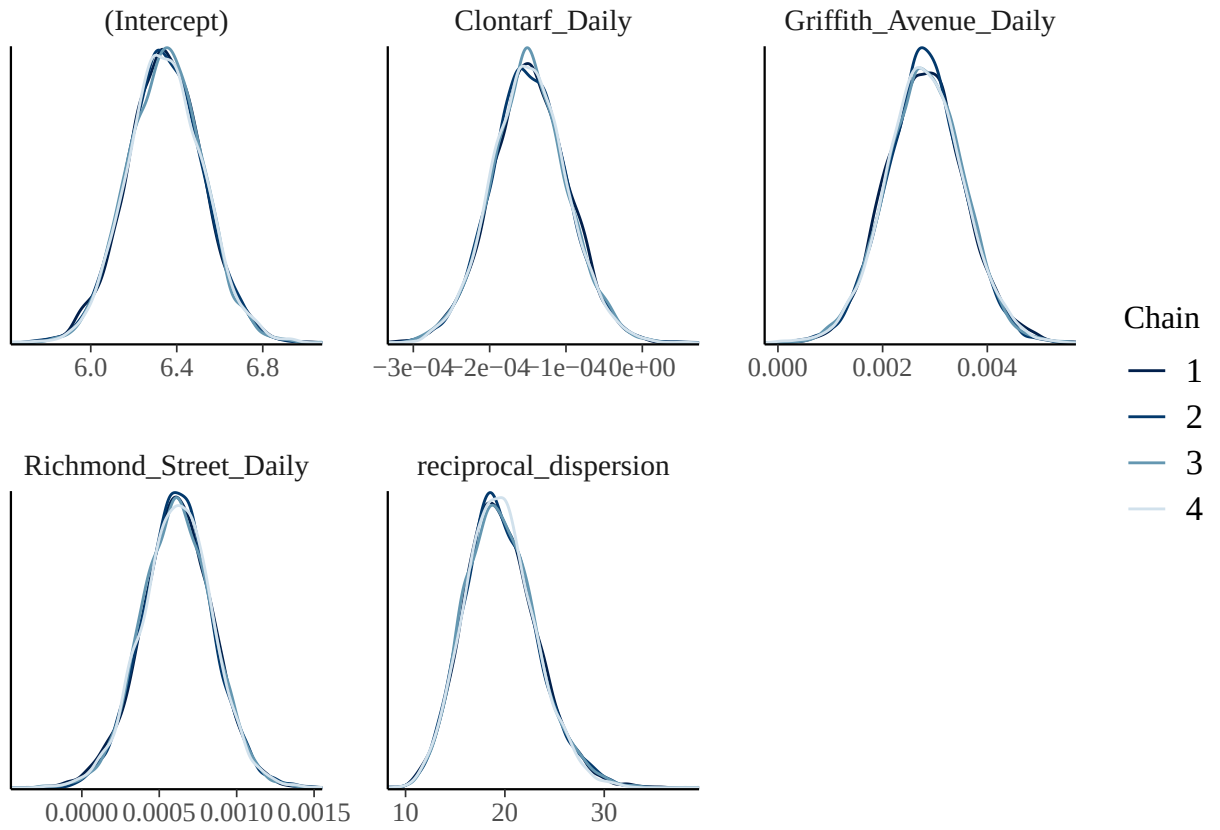
We observe that Richmond Street and Griffith Avenue are positively correlated with Grove Road cyclist volumes, with Griffith Avenue having the largest positive coefficient. All explanatory variables are significant given that none of their respective confidence intervals contain zero.

```
mcmc_trace(bikes_model_negbin)
```



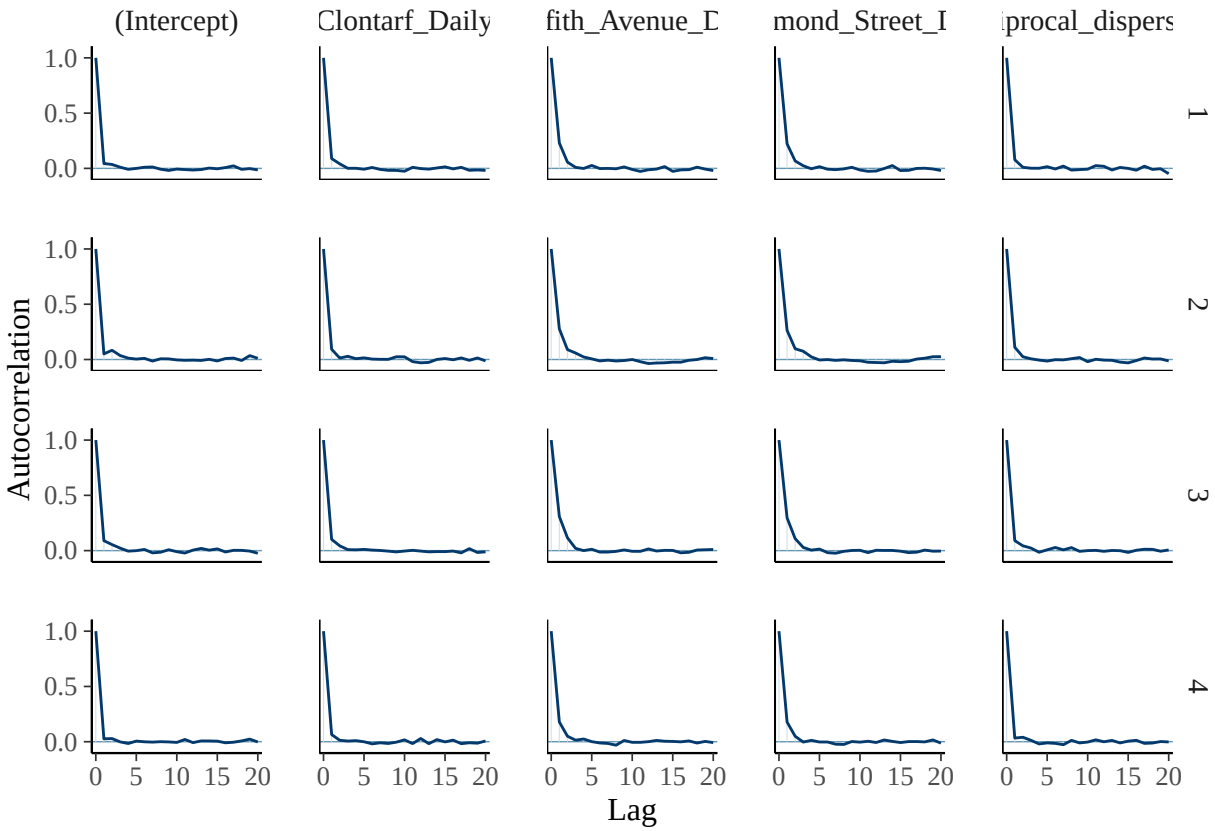
I have plotted the MCMC trace plots of the samples for each chain and there doesn't look to be any issues. There are no apparent anomalies or serial correlation between draws and the chain seems to explore the sample space many times.

```
mcmc_dens_overlay(bikes_model_negbin)
```



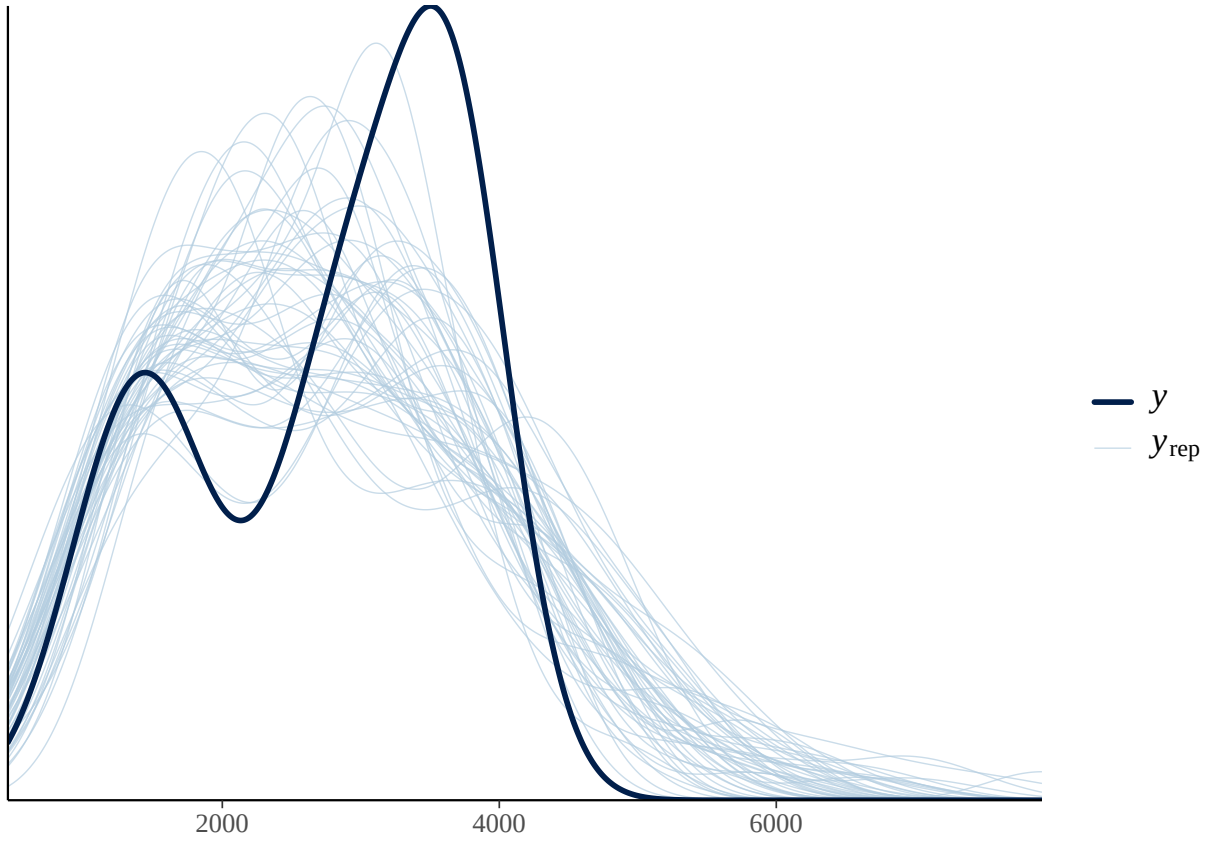
Above I plot the densities for each variable and chain overlayed on each other to ensure each chain is producing similar densities for the variables. Per the above plot we can see that this is the case and all looks to be in order.

```
mcmc_acf(bikes_model_negbin)
```



Above, I have plotted the ACF plots for each variable and chain. Again there are no apparent issues, there is significant (expected) autocorrelation at short lags but these tend to zero very quickly for all chains and variables.

```
pp_check(bikes_model_negbin)
```



Above I plot the observed density of Y (Grove Road cyclists) against 50 simulated densities from the posterior probability distribution. We can see that the simulated densities are similar however many are failing to capture the multimodal nature of the observed dataset, suggesting we are missing some explantory variable or not fully capturing the variance within the observed data here.

```
prediction_summary(model = bikes_model_negbin, data = bikes_q2)
```

```
##        mae mae_scaled within_50 within_95
## 1 224.2868  0.3811889 0.7868852         1
```

Above, I plot the model prediction summary to assess goodness of fit of the model. We see that the mean absolute error of the model is 224 and the scaled standard mean absolute error is 0.38 standard deviations.

Question 3

Is there a relationship between the number of cyclists and the weather?

As we are again dealing with counts of cyclists in this question, I will propose to use either a Poisson or Negative-Binomial GLM data model. For consistency and comparative purposes I will again use the daily Grove Road cyclist volume as my target variable and the available average daily weather variables as my explanatory variables.

I propose using the following weather-based variables as explanatory variables for this question; average daily rain (mm), average daily temp ("Ŏ0B0C"), average daily wdsp (kt) and average daily visibility (m). I have chosen not to include the cloud amount (okta) as it's a categorical variable which does not seem to have any significant relationship with the volume of cyclists on Grove Road.

```r
# Aggregating cyclist counts to daily figures and weather data to average daily values
bikes_q3 <- bikes %>%
  select(Grove_Road,Date,rain,temp,wdsp,vis) %>%
  group_by(Date) %>%
  mutate(Grove_Road_Daily = sum(Grove_Road),
         rain_daily_avg = mean(rain),
         temp_daily_avg = mean(temp),
         wdsp_daily_avg = mean(wdsp),
         vis_daily_avg = mean(vis)) %>%
  ungroup() %>%
  select(-Grove_Road,-rain,-temp,-wdsp,-vis) %>%
  unique()
```

Firstly I observe the distributions of each of the daily average weather data help determine normality of variables and identify suitable priors.

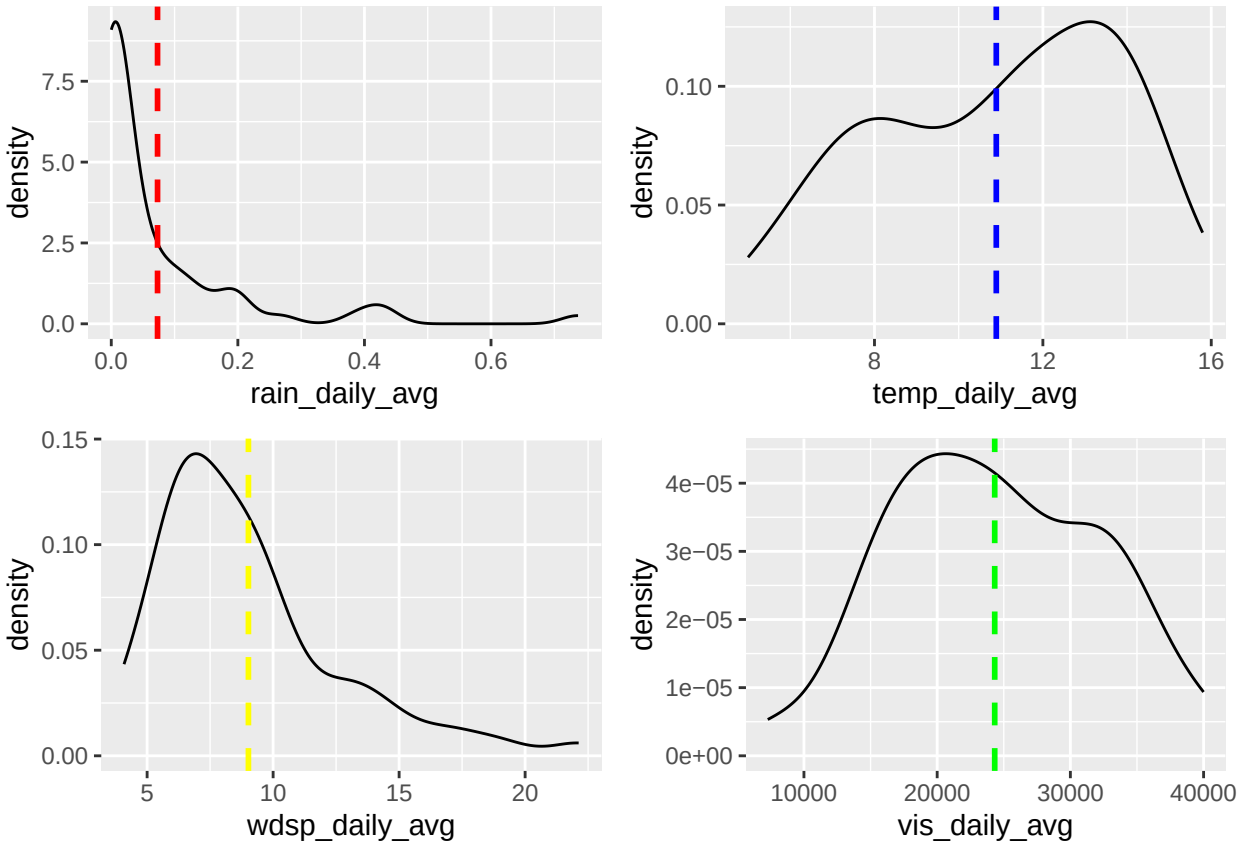
```r
# Checking distributions of target and response variables
p1 <- ggplot(bikes_q3,aes(x=rain_daily_avg)) +
        geom_density() +
        geom_vline(aes(xintercept=mean(rain_daily_avg)),
            color="red", linetype="dashed", linewidth=1)

p2 <- ggplot(bikes_q3,aes(x=temp_daily_avg)) +
        geom_density() +
        geom_vline(aes(xintercept=mean(temp_daily_avg)),
            color="blue", linetype="dashed", linewidth=1)

p3 <- ggplot(bikes_q3,aes(x=wdsp_daily_avg)) +
        geom_density() +
        geom_vline(aes(xintercept=mean(wdsp_daily_avg)),
            color="yellow", linetype="dashed", linewidth=1)

p4 <- ggplot(bikes_q3,aes(x=vis_daily_avg)) +
        geom_density() +
        geom_vline(aes(xintercept=mean(vis_daily_avg)),
            color="green", linetype="dashed", linewidth=1)

grid.arrange(p1,p2,p3,p4,nrow = 2, ncol = 2)
```



As we can see from the distribution plots above, wdsp, temp, and vis are roughly normal in distribution however rain is heavily right-skewed.

I will use the same prior for the intercept as in question 2 as I am using the same target variable ($N \sim (8, 2)$). I use the default weakly-informative priors from rstanarm for the prior distributions on the beta coefficients.

```
# Estimating posterior using rstanarm
gr_weather_model_negbin_prior <- stan_glm(Grove_Road_Daily ~ rain_daily_avg + temp_daily_avg +
                                            wdsp_daily_avg + vis_daily_avg,
                                          data = bikes_q3,
                                          family = neg_binomial_2,
                                          #prior for beta_0
                                          prior_intercept = normal(8,2),
                                          #tune remaining priors
                                          # prior = normal(0,10,autoscale = TRUE),
                                          # prior = my_priors,
                                          #tune prior for r
                                          prior_aux = exponential(1, autoscale = TRUE),
                                          chains = 4,
                                          iter = 5000*2,
                                          seed = 1759,
                                          #warmup = 0,
                                          refresh = 0,
                                          prior_PD = TRUE)
```

```r
# Estimate posterior using rstanarm
gr_weather_model_negbin_prior$prior.info
```

```
## $prior
## $prior$dist
## [1] "normal"
##
## $prior$location
## [1] 0 0 0 0
##
## $prior$scale
## [1] 2.5 2.5 2.5 2.5
##
## $prior$adjusted_scale
## [1] 1.863133e+01 8.693455e-01 6.742955e-01 3.217261e-04
##
## $prior$df
## NULL
##
##
## $prior_intercept
## $prior_intercept$dist
## [1] "normal"
##
## $prior_intercept$location
## [1] 8
##
## $prior_intercept$scale
## [1] 2
##
## $prior_intercept$adjusted_scale
## NULL
##
## $prior_intercept$df
## NULL
##
##
## $prior_aux
## $prior_aux$dist
## [1] "exponential"
##
## $prior_aux$location
## NULL
##
## $prior_aux$scale
## NULL
##
## $prior_aux$adjusted_scale
## NULL
##
## $prior_aux$df
## NULL
##
```

```
## $prior_aux$rate
## [1] 1
##
## $prior_aux$aux_name
## [1] "reciprocal_dispersion"
```

Based on the above output stan_glm gives the following priors:

$\beta_1 \sim Normal(0, 18.63^2)$ $\beta_2 \sim Normal(0, 0.87^2)$ $\beta_3 \sim Normal(0, 0.67^2)$ $\beta_4 \sim Normal(0, 0.0003^2)$ $r \sim Exp(1)$

```
# Simulating the posterior
gr_weather_model_negbin <- update(gr_weather_model_negbin_prior, prior_PD = FALSE)
tidy(gr_weather_model_negbin, conf.int = TRUE, conf.level = 0.95)
```

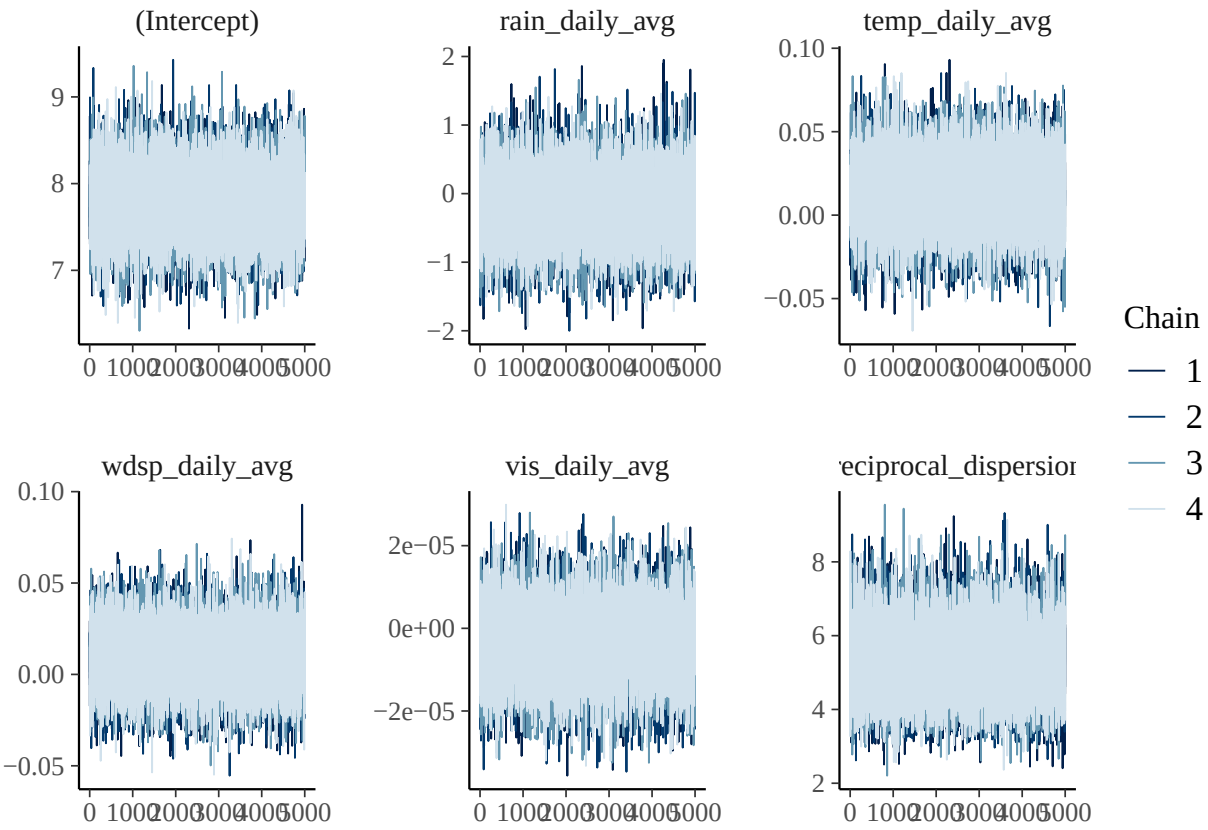
```
## # A tibble: 5 x 5
##   term              estimate  std.error   conf.low conf.high
##   <chr>                <dbl>      <dbl>      <dbl>     <dbl>
## 1 (Intercept)     7.79       0.376       7.06      8.56
## 2 rain_daily_avg -0.238      0.481      -1.13      0.764
## 3 temp_daily_avg  0.0140     0.0205     -0.0267    0.0538
## 4 wdsp_daily_avg  0.00914    0.0166     -0.0229    0.0421
## 5 vis_daily_avg  -0.00000352 0.00000850 -0.0000208 0.0000132
```

Above we observe the Bayesian regression output for the weather model. We can see there is a large intercept coefficient of 7.79 and is significant given that the 95% interval is far from (and does not contain) zero.

We observe negative coefficients for rain and visibility variables. The negative coefficient for rainfall in mm makes intuitive sense given that many people would not cycle in the rain however we would expect the greater the visibility the higher the volume of cyclists on Grove Road. We also notice that each of the explantary variables confidence intervals contain zero and so we cannot rule out the possibility they are zero.

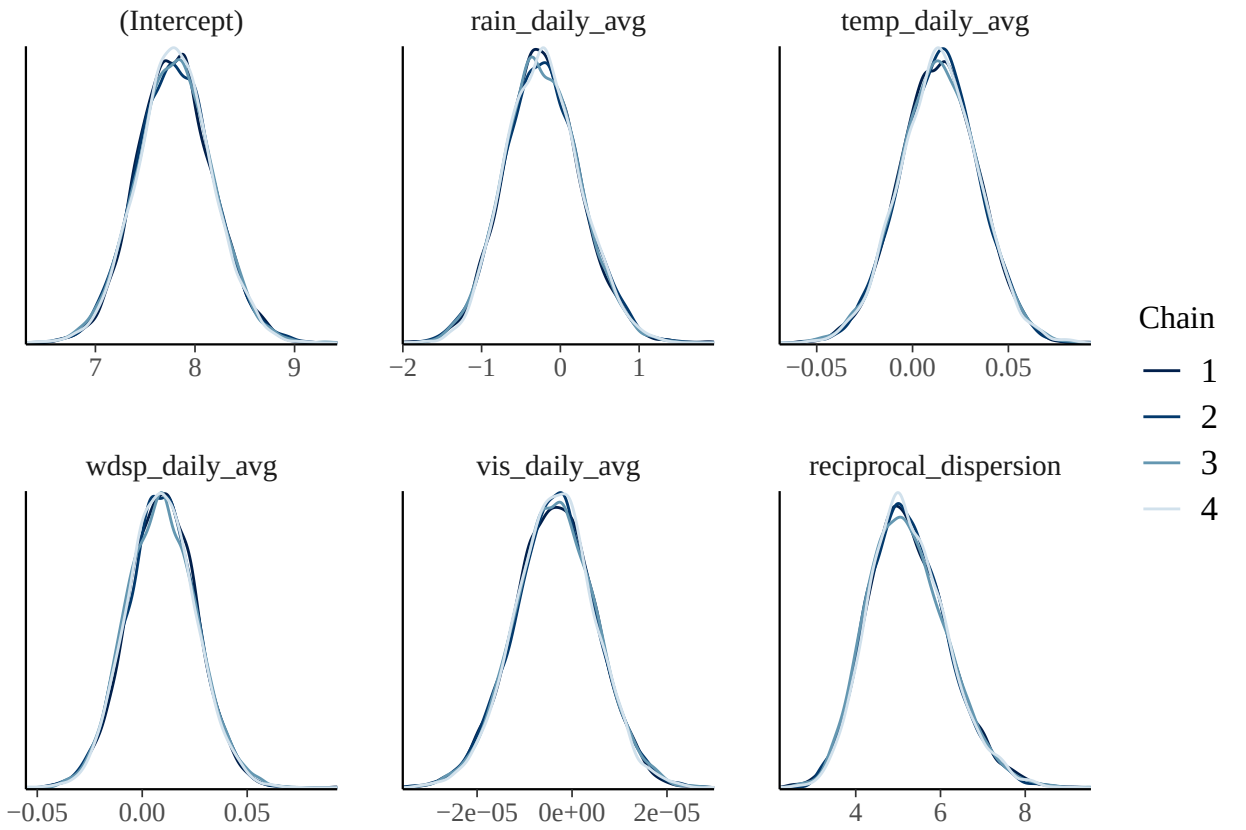
We observe that temp and wdsp are positively correlated with Grove Road cyclist volumes, intuitively the temp makes sense as you would expect less people would cycle when it's colder particularly if it's cold enough for ice to form on roads as this makes cycling much more dangerous. We do not expect high winds to be associated with higher volumes of cyclists however.

```
mcmc_trace(gr_weather_model_negbin)
```



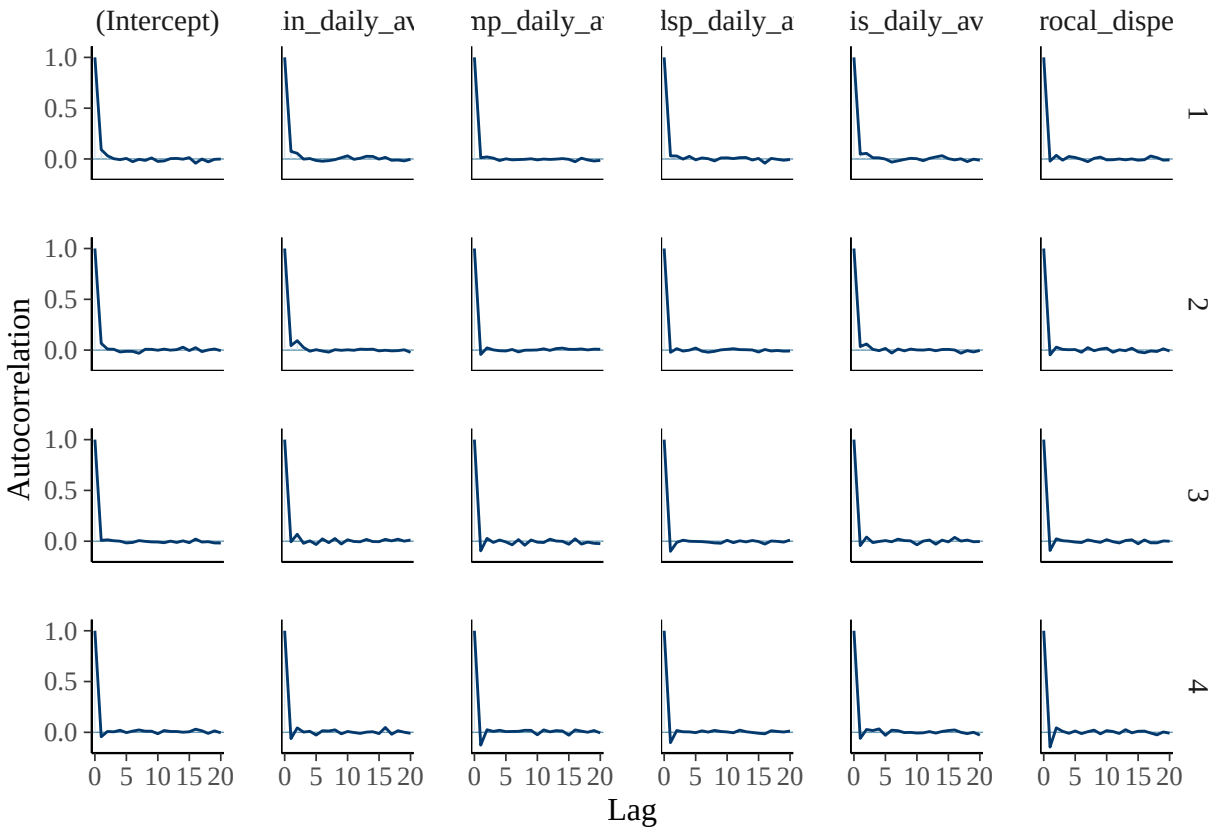
The MCMC trace plots all look good here and draws seem to be non-correlated and coming from entire sample space.

```
mcmc_dens_overlay(gr_weather_model_negbin)
```



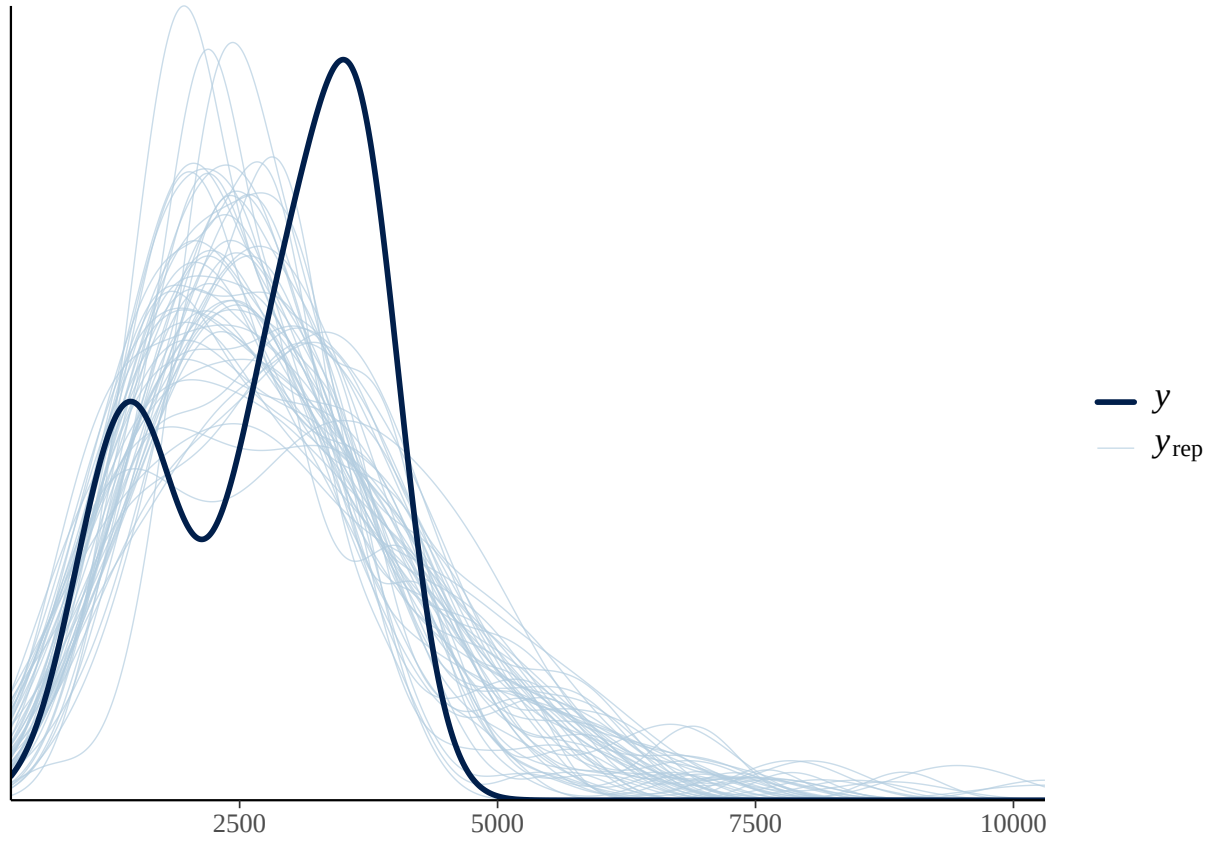
Densities of variables across the 4 chains all look similar here.

```
mcmc_acf(gr_weather_model_negbin)
```



ACF plots all look good and do not signal any issues with the MCMC draws.

```
pp_check(gr_weather_model_negbin)
```



Observing the above plot of the 50 simulated densities from the posterior against the observed Y density we see that we are still not fully capturing the multi-modality of the observed Y with the posterior and perhaps are missing a layer of complexity / explanatory variable.

```
prediction_summary(model = gr_weather_model_negbin, data = bikes_q3)
```

```
##        mae mae_scaled within_50 within_95
## 1 764.2147  0.5885767  0.442623         1
```

I have produced the prediction summary output below, we observe a much higher mae and scaled mae than the model in question 2 indicating a poorer model fit.

Question 4

Based on this dataset, choose another scientific question and perform an appropriate Bayesian GLM.

Based on this dataset, I am interested if there is a relationship between the number of cyclists and whether it is a weekend or weekday.

In order to attempt to answer this question I will again use the daily volume of cyclists on Grove Road as my target variable and I will create 2 new binary variables namely; *weekend* which will be equal to 1 if the day is Saturday or Sunday and 0 otherwise and *mon_or_fri* which will be equal to 1 if the day is a Monday or Friday and 0 otherwise.

The motivation for including a Monday/Friday variable stems from the introduction of work from home practices in many offices in Dublin with many people choosing to work from home on a Monday/Friday and therefore not have to commute to the office.

```
bikes_q4 <- bikes %>%
  group_by(Date) %>%
  mutate(Grove_Road_Daily = sum(Grove_Road),
         Weekend = case_when(Day %in% c("Sat","Sun") ~ 1,
                             TRUE ~ 0),
         Mon_or_Fri = case_when(Day %in% c("Mon","Fri") ~ 1,
                                TRUE ~ 0)) %>%
  ungroup()
```

I will use the same prior for the intercept as in question 2 as I am using the same target variable ($N \sim (8, 2)$). I use the default weakly-informative priors from rstanarm for the prior distributions on the beta coefficients. I will also again be using a negative-binomial model to fit to the data.

```
# Estimating posterior using rstanarm
gr_day_model_negbin_prior <- stan_glm(Grove_Road_Daily ~ Weekend + Mon_or_Fri,
                                      data = bikes_q4,
                                      family = neg_binomial_2,
                                      #prior for beta_0
                                      prior_intercept = normal(8,2),
                                      #tune remaining priors
                                      prior = normal(0,2.5,autoscale = TRUE),
                                      #tune prior for r
                                      prior_aux = exponential(1, autoscale = TRUE),
                                      chains = 4,
                                      iter = 5000*2,
                                      seed = 1759,
                                      #warmup = 0,
                                      refresh = 0,
                                      prior_PD = TRUE)
```

```
# Estimate posterior using rstanarm
gr_day_model_negbin_prior$prior.info
```

```
## $prior
## $prior$dist
## [1] "normal"
##
## $prior$location
## [1] 0 0
##
## $prior$scale
## [1] 2.5 2.5
##
## $prior$adjusted_scale
## [1] 5.681398 5.479627
##
## $prior$df
## NULL
##
```

```
##
## $prior_intercept
## $prior_intercept$dist
## [1] "normal"
##
## $prior_intercept$location
## [1] 8
##
## $prior_intercept$scale
## [1] 2
##
## $prior_intercept$adjusted_scale
## NULL
##
## $prior_intercept$df
## NULL
##
##
## $prior_aux
## $prior_aux$dist
## [1] "exponential"
##
## $prior_aux$location
## NULL
##
## $prior_aux$scale
## NULL
##
## $prior_aux$adjusted_scale
## NULL
##
## $prior_aux$df
## NULL
##
## $prior_aux$rate
## [1] 1
##
## $prior_aux$aux_name
## [1] "reciprocal_dispersion"
```

Based on the above output stan_glm gives the following priors:

$\beta_1 \sim Normal(0, 5.68^2)$ $\beta_2 \sim Normal(0, 5.48^2)$ $r \sim Exp(1)$

```
# Simulating the posterior
gr_day_model_negbin <- update(gr_day_model_negbin_prior, prior_PD = FALSE)
tidy(gr_day_model_negbin, conf.int = TRUE, conf.level = 0.95)
```

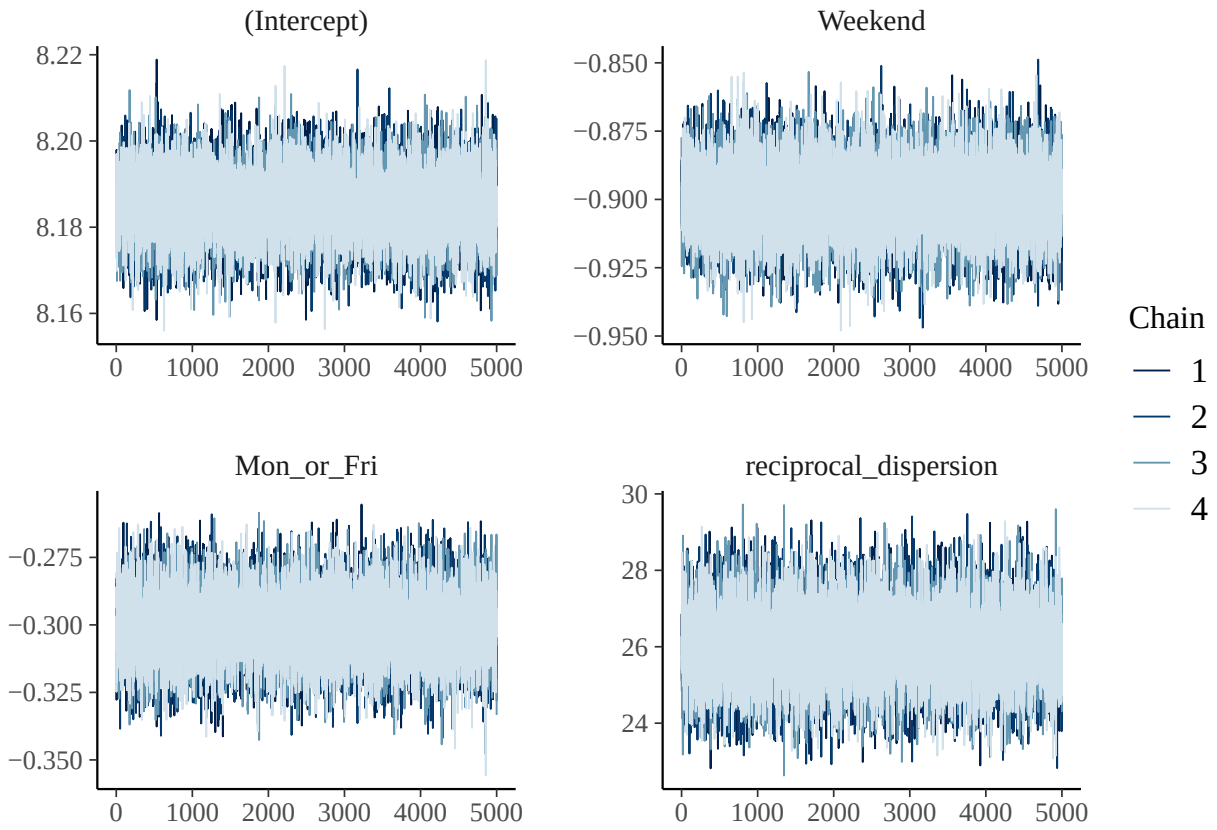
```
## # A tibble: 3 x 5
##   term        estimate std.error conf.low conf.high
##   <chr>          <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)    8.19    0.00774    8.17      8.20
## 2 Weekend       -0.899   0.0126    -0.925    -0.874
## 3 Mon_or_Fri    -0.300   0.0122    -0.324    -0.276
```

Above we observe the output of the Bayesian negative binomial regression model. We note that all coefficients are significant given their respective confidence intervals do not contain zero. Again we have a large positive coefficient for the intercept which makes sense given we are dealing with positive counts of cyclists.

We note that both the Weekend and Mon_or_Fri variables are negatively correlated with the number of cyclists on Grove Road which intuitively makes sense. The Grove Road is used extensively by commuters and as such the weekend would be negatively correlated with cycling traffic as there is less people commuting.

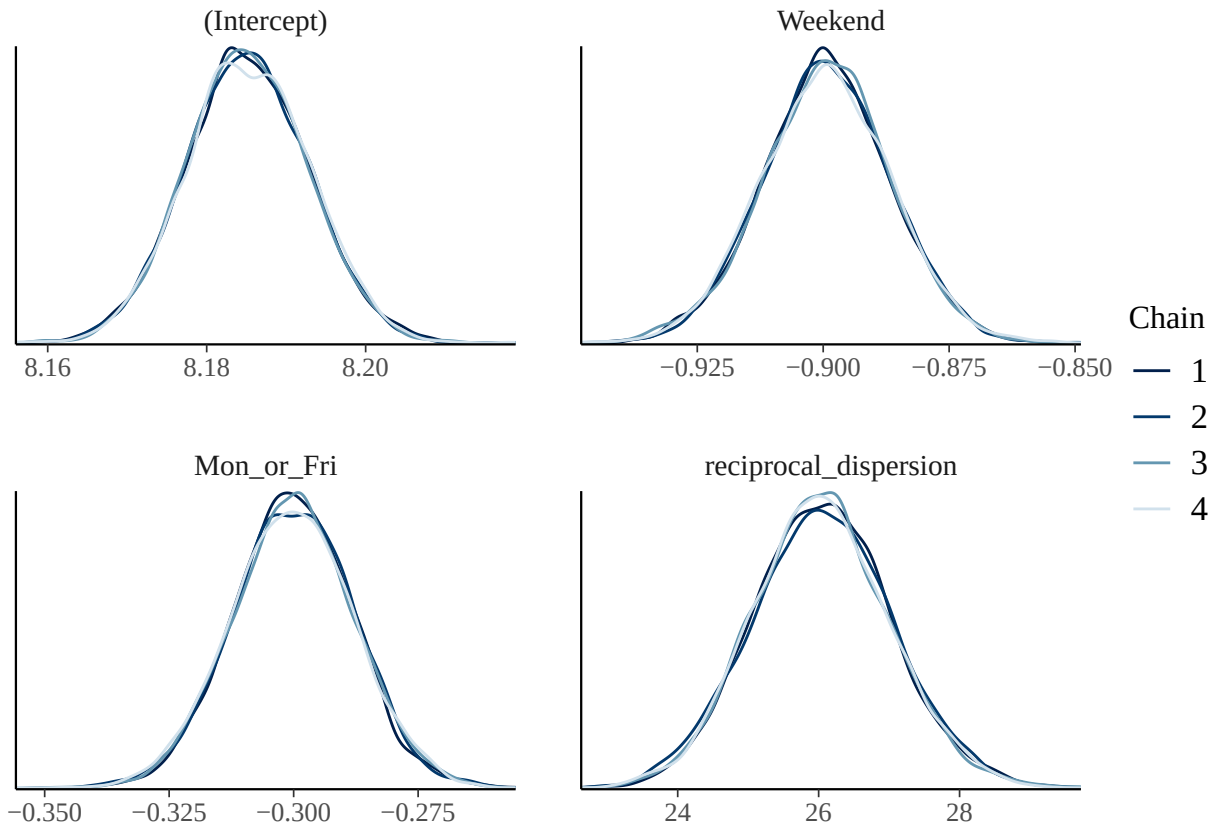
It is interesting to see that the Mon_or_Fri variable is also negatively correlated but not to the magnitude as the Weekend variable which possibly indicates the reduction of cycling traffic due to work-from-home policies in the city but further analysis would be required to determine the causality here.

```
mcmc_trace(gr_day_model_negbin)
```



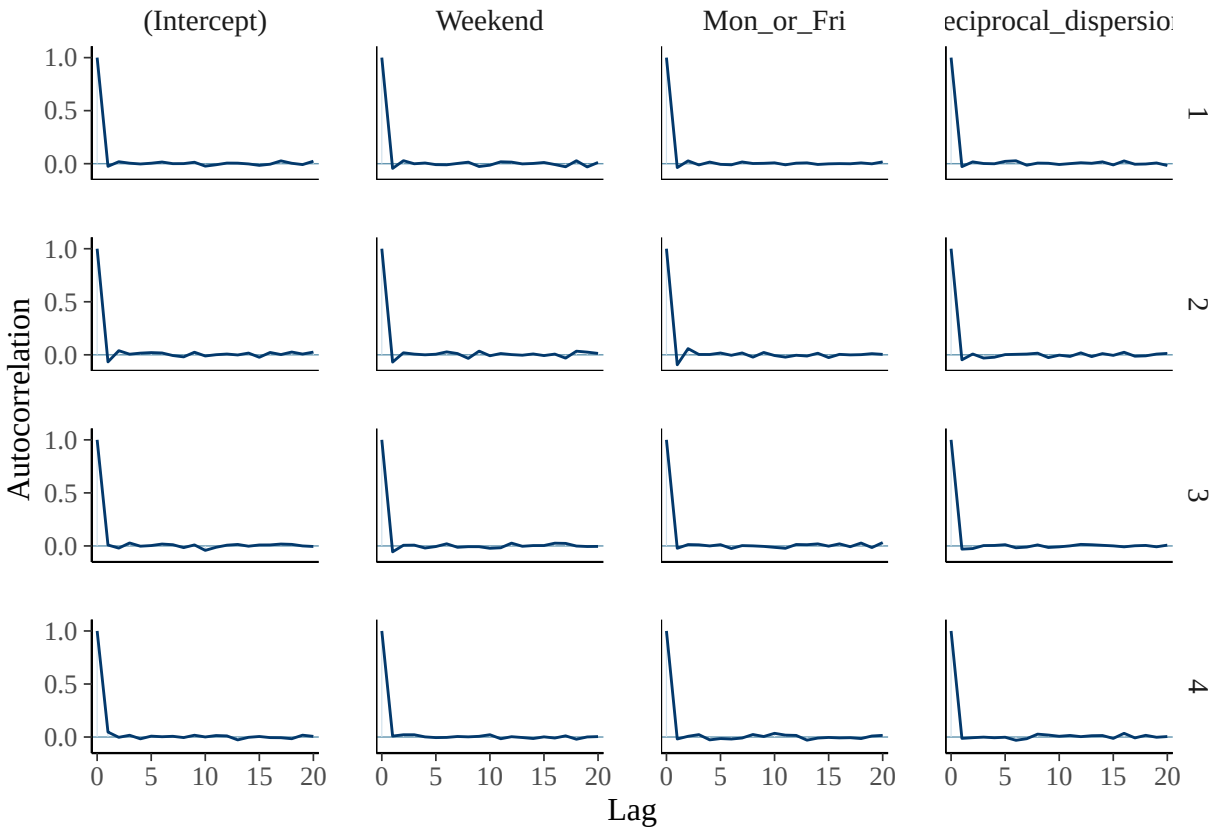
All trace plots look ok.

```
mcmc_dens_overlay(gr_day_model_negbin)
```



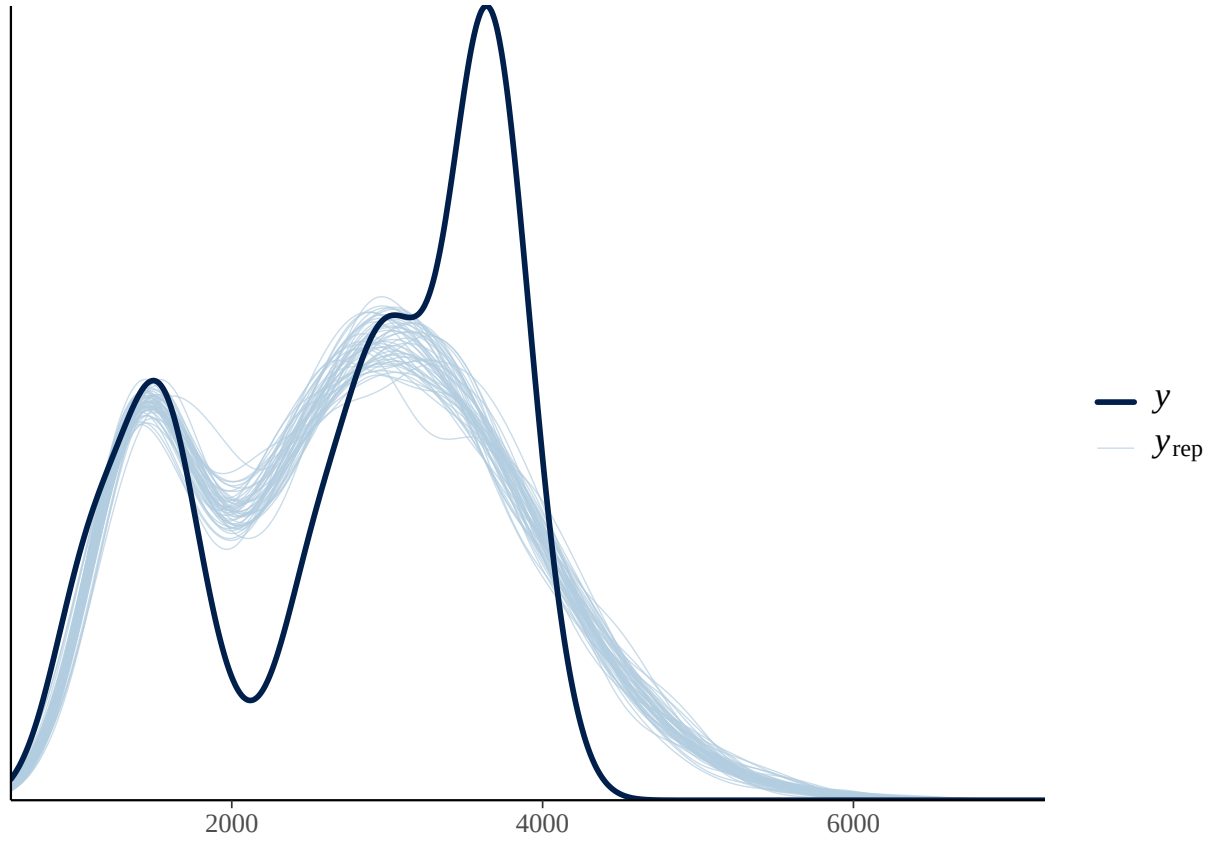
All densities look similar for each chain for each variable.

```
mcmc_acf(gr_day_model_negbin)
```



ACF plots also look good, no issues arising from any of the diagnostic plots.

```
pp_check(gr_day_model_negbin)
```



Observing the above observed Y density against the 50 simulated densities from the posterior we can see this is much closer to capturing the observed distribution than the previous 2 models. There is still room for improvement however as the peak and trough of the observed data is not quite being fully captured by the simulations from the posterior.

```
prediction_summary(model = gr_day_model_negbin, data = bikes_q4)
```

```
##        mae mae_scaled within_50 within_95
## 1 178.2751  0.3184493  0.704235 0.9508197
```

Comment on the above plot will be made below as part of Q5.

Question 5

What's the best model between the one's proposed?

In order to answer the above question I will compare the models presented in q2-q4. As in each of these models I have used the same target variable, namely the daily volume of cyclists passing through the Grove Road totem, it will be possible to compare these models using mean absolute error (mae) or scaled mean absolute error (mae_scaled).

Observing the prediction summary outputs from the bayesrules package, we can see that the model from q4 (weekend / mon_or_fri) model has the lowest mean absolute error and scaled mean absolute error of the 3 models. This means that typically, the difference between observed y (# cyclists using Grove Road daily) and the posterior predictive means is lowest for this model. The next best model is the model used in question 2 with an mae of 224 and a scaled mae of 0.34 standard deviations. The worst fitting model to

the observed cyclist data from Grove Road is the weather model which has an mae of 764 and scaled mae of 0.59 standard deviations which reflects the low correlation between the weather variables and number of cyclists.