451

Jeremy Liu,Rachel Rubanguka Hoops,Xieyao Yin,Casey Lee

2024-12-02

```r
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.1
```

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

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

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

```r
library(rstan)
```

```
## Warning: package 'rstan' was built under R version 4.4.1
```

```
## Loading required package: StanHeaders
```

```
## Warning: package 'StanHeaders' was built under R version 4.4.1
```

```
##
## rstan version 2.32.6 (Stan version 2.32.2)
```

```
## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)
## For within-chain threading using `reduce_sum()` or `map_rect()` Stan functions,
## change `threads_per_chain` option:
## rstan_options(threads_per_chain = 1)
```

```
## Do not specify '-march=native' in 'LOCAL_CPPFLAGS' or a Makevars file
```

```r
library(bayesplot)
```

```
## Warning: package 'bayesplot' was built under R version 4.4.1

## This is bayesplot version 1.11.1

## - Online documentation and vignettes at mc-stan.org/bayesplot

## - bayesplot theme set to bayesplot::theme_default()

##     * Does _not_ affect other ggplot2 plots

##     * See ?bayesplot_theme_set for details on theme setting
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.2

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.4.1

## Loading required package: lattice
```

```r
library(posterior)
```

```
## Warning: package 'posterior' was built under R version 4.4.1

## This is posterior version 1.6.0

##
## Attaching package: 'posterior'

## The following object is masked from 'package:bayesplot':
##
##     rhat

## The following objects are masked from 'package:rstan':
##
##     ess_bulk, ess_tail

## The following objects are masked from 'package:stats':
##
##     mad, sd, var

## The following objects are masked from 'package:base':
##
##     %in%, match
```

```r
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.1
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:rstan':
##
##     extract
```

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.4.1
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.4.2
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```r
library(brms)
```

```
## Warning: package 'brms' was built under R version 4.4.1
```

```
## Loading required package: Rcpp
```

```
## Warning: package 'Rcpp' was built under R version 4.4.1
```

```
## Loading 'brms' package (version 2.21.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').
```

```
##
## Attaching package: 'brms'
```

```
## The following object is masked from 'package:bayesplot':
##
##     rhat
```

```
## The following object is masked from 'package:rstan':
##
##     loo


## The following object is masked from 'package:stats':
##
##     ar
```

```r
set.seed(123)
```

```r
#knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
#library(dplyr)
#library(tidyr)

data = read.csv("US_Accidents_March23_sampled_500k.csv")

head(data)
```

```
##          ID  Source Severity                          Start_Time
## 1 A-2047758 Source2        2                 2019-06-12 10:10:56
## 2 A-4694324 Source1        2 2022-12-03 23:37:14.000000000
## 3 A-5006183 Source1        2 2022-08-20 13:13:00.000000000
## 4 A-4237356 Source1        2                 2022-02-21 17:43:04
## 5 A-6690583 Source1        2                 2020-12-04 01:46:00
## 6 A-1101469 Source2        2                 2021-03-29 07:03:58
##                         End_Time Start_Lat Start_Lng  End_Lat    End_Lng
## 1            2019-06-12 10:55:58   30.64121  -91.15348       NA         NA
## 2 2022-12-04 01:56:53.000000000   38.99056  -77.39907 38.99004  -77.39828
## 3 2022-08-20 15:22:45.000000000   34.66119 -120.49282 34.66119 -120.49244
## 4            2022-02-21 19:43:23   43.68059  -92.99332 43.68057  -92.97222
## 5            2020-12-04 04:13:09   35.39548 -118.98518 35.39548 -118.98600
## 6            2021-03-29 08:51:01   42.53208  -70.94427       NA         NA
##   Distance.mi.
## 1        0.000
## 2        0.056
## 3        0.022
## 4        1.054
## 5        0.046
## 6        0.000
##                                                         Description
## 1         Accident on LA-19 Baker-Zachary Hwy at Lower Zachary Rd.
## 2 Incident on FOREST RIDGE DR near PEPPERIDGE PL Drive with caution.
## 3      Accident on W Central Ave from Floradale Ave to Western Ave.
## 4           Incident on I-90 EB near REST AREA Drive with caution.
## 5          RP ADV THEY LOCATED SUSP VEH OF 20002 - 726 CRAWFORD
## 6                          Accident on Forest St at Lowell St.
##            Street       City         County State   Zipcode Country
## 1     Highway 19    Zachary East Baton Rouge    LA 70791-4610      US
## 2 Forest Ridge Dr   Sterling        Loudoun    VA 20164-2813      US
## 3   Floradale Ave     Lompoc  Santa Barbara    CA      93436      US
## 4      14th St NW     Austin          Mower    MN      55912      US
## 5      River Blvd Bakersfield           Kern    CA 93305-2649      US
```

```
## 6          Lowell St       Peabody              Essex    MA 01960-4275        US
##       Timezone Airport_Code   Weather_Timestamp Temperature.F. Wind_Chill.F.
## 1 US/Central         KBTR 2019-06-12 09:53:00             77            77
## 2 US/Eastern         KIAD 2022-12-03 23:52:00             45            43
## 3 US/Pacific         KLPC 2022-08-20 12:56:00             68            68
## 4 US/Central         KAUM 2022-02-21 17:35:00             27            15
## 5 US/Pacific         KBFL 2020-12-04 01:54:00             42            42
## 6 US/Eastern         KBVY 2021-03-29 06:53:00             42            35
##   Humidity... Pressure.in. Visibility.mi. Wind_Direction Wind_Speed.mph.
## 1          62        29.92             10             NW               5
## 2          48        29.91             10              W               5
## 3          73        29.79             10              W              13
## 4          86        28.49             10            ENE              15
## 5          34        29.77             10           CALM               0
## 6          58        29.37             10              W              13
##   Precipitation.in. Weather_Condition Amenity  Bump Crossing Give_Way Junction
## 1                 0              Fair   False False    False    False    False
## 2                 0              Fair   False False    False    False    False
## 3                 0              Fair   False False    False    False    False
## 4                 0        Wintry Mix   False False    False    False    False
## 5                 0              Fair   False False    False    False    False
## 6                 0              Fair   False False    False    False    False
##   No_Exit Railway Roundabout Station  Stop Traffic_Calming Traffic_Signal
## 1   False   False      False   False False           False           True
## 2   False   False      False   False False           False          False
## 3   False   False      False   False False           False           True
## 4   False   False      False   False False           False          False
## 5   False   False      False   False False           False          False
## 6   False   False      False   False False           False           True
##   Turning_Loop Sunrise_Sunset Civil_Twilight Nautical_Twilight
## 1        False            Day            Day               Day
## 2        False          Night          Night             Night
## 3        False            Day            Day               Day
## 4        False            Day            Day               Day
## 5        False          Night          Night             Night
## 6        False            Day            Day               Day
##   Astronomical_Twilight
## 1                   Day
## 2                 Night
## 3                   Day
## 4                   Day
## 5                 Night
## 6                   Day
```

```
colnames(data)
```

```
##  [1] "ID"                  "Source"              "Severity"
##  [4] "Start_Time"          "End_Time"            "Start_Lat"
##  [7] "Start_Lng"           "End_Lat"             "End_Lng"
## [10] "Distance.mi."        "Description"         "Street"
## [13] "City"                "County"              "State"
## [16] "Zipcode"             "Country"             "Timezone"
## [19] "Airport_Code"        "Weather_Timestamp"   "Temperature.F."
## [22] "Wind_Chill.F."       "Humidity..."         "Pressure.in."
```

```
## [25] "Visibility.mi."       "Wind_Direction"       "Wind_Speed.mph."
## [28] "Precipitation.in."     "Weather_Condition"    "Amenity"
## [31] "Bump"                  "Crossing"             "Give_Way"
## [34] "Junction"              "No_Exit"              "Railway"
## [37] "Roundabout"            "Station"              "Stop"
## [40] "Traffic_Calming"       "Traffic_Signal"       "Turning_Loop"
## [43] "Sunrise_Sunset"        "Civil_Twilight"       "Nautical_Twilight"
## [46] "Astronomical_Twilight"
```

```r
cleaned_data <- data %>% select(
  Severity, Distance.mi., Description, Start_Lat, Start_Lng, City, County, State,
  Start_Time, End_Time, Timezone, Temperature.F., Wind_Chill.F., Humidity...,
  Pressure.in., Visibility.mi., Wind_Direction, Wind_Speed.mph., Precipitation.in.,
  Weather_Condition, Amenity, Bump, Crossing, Junction, No_Exit,
  Traffic_Calming, Traffic_Signal
)

cleaned_data <- cleaned_data %>%
  select(
    which(colSums(is.na(.)) / nrow(.) < 0.5)
  )

cleaned_data <- cleaned_data %>% drop_na(Severity, Start_Time, Start_Lat, Start_Lng)

cleaned_data <- cleaned_data %>%
  mutate(
    Start_Time = as.POSIXct(Start_Time, format = "%Y-%m-%d %H:%M:%S"),
    End_Time = as.POSIXct(End_Time, format = "%Y-%m-%d %H:%M:%S")
  )

categorical_cols <- c("City", "County", "State", "Timezone", "Weather_Condition",
                      "Wind_Direction", "Amenity", "Bump", "Crossing", "Junction",
                      "No_Exit", "Traffic_Calming", "Traffic_Signal")
cleaned_data <- cleaned_data %>% mutate(across(all_of(categorical_cols), as.factor))




cleaned_data <- cleaned_data %>%
  mutate(Duration = as.numeric(difftime(End_Time, Start_Time, units = "mins")))

cleaned_data <- cleaned_data %>%
  mutate(Time_of_Day = case_when(
    format(Start_Time, "%H") %in% c(6:11) ~ "Morning",
    format(Start_Time, "%H") %in% c(12:17) ~ "Afternoon",
    format(Start_Time, "%H") %in% c(18:21) ~ "Evening",
    TRUE ~ "Night"
  ))

cleaned_data <- na.omit(cleaned_data)
print(colnames(cleaned_data))
```

```
##  [1] "Severity"         "Distance.mi."      "Description"
##  [4] "Start_Lat"        "Start_Lng"         "City"
##  [7] "County"           "State"             "Start_Time"
```

```
## [10] "End_Time"         "Timezone"           "Temperature.F."
## [13] "Wind_Chill.F."    "Humidity..."        "Pressure.in."
## [16] "Visibility.mi."   "Wind_Direction"     "Wind_Speed.mph."
## [19] "Precipitation.in." "Weather_Condition" "Amenity"
## [22] "Bump"             "Crossing"           "Junction"
## [25] "No_Exit"          "Traffic_Calming"    "Traffic_Signal"
## [28] "Duration"         "Time_of_Day"
```

```r
# save to csv
write.csv(cleaned_data, "cleaned_data.csv", row.names = FALSE)

cleaned_data <- read.csv("cleaned_data.csv")

sampled_data <- cleaned_data %>%
  group_by(Severity) %>%
  slice_sample(n = 400) %>%
  ungroup()

head(sampled_data)
```

```
## # A tibble: 6 x 29
##   Severity Distance.mi. Description     Start_Lat Start_Lng City  County State
##      <int>        <dbl> <chr>               <dbl>     <dbl> <chr> <chr>  <chr>
## 1        1            0 Crash on FL-589-~    28.1     -82.6 Tampa Hills~ FL
## 2        1            0 Crash on FL-64 a~    27.5     -82.5 Brad~ Manat~ FL
## 3        1            0 Crash on Mudd Av~    30.2     -92.0 Lafa~ Lafay~ LA
## 4        1            0 At Jupiter Blvd ~    28.0     -80.7 Palm~ Breva~ FL
## 5        1            0 Crash on US-290 ~    30.2     -98.0 Aust~ Travis TX
## 6        1            0 At Esplanade Ave~    30.0     -90.1 New ~ Orlea~ LA
## # i 21 more variables: Start_Time <chr>, End_Time <chr>, Timezone <chr>,
## #   Temperature.F. <dbl>, Wind_Chill.F. <dbl>, Humidity... <int>,
## #   Pressure.in. <dbl>, Visibility.mi. <dbl>, Wind_Direction <chr>,
## #   Wind_Speed.mph. <dbl>, Precipitation.in. <dbl>, Weather_Condition <chr>,
## #   Amenity <chr>, Bump <chr>, Crossing <chr>, Junction <chr>, No_Exit <chr>,
## #   Traffic_Calming <chr>, Traffic_Signal <chr>, Duration <dbl>,
## #   Time_of_Day <chr>
```

```r
write.csv(sampled_data, "sampled_data4.csv", row.names = FALSE)
```

main

```r
cleaned_data <- read.csv("sampled_data4.csv")
cleaned_data <- cleaned_data %>%
  select(Severity, Crossing, Amenity, Traffic_Signal, Temperature.F., Junction,
         Distance.mi., Wind_Speed.mph., Precipitation.in.)


# CONVERT INT
binary_columns <- c("Amenity", "Traffic_Signal", "Junction", "Crossing")
cleaned_data[binary_columns] <- lapply(cleaned_data[binary_columns], function(x) as.numeric(as.logical(
cleaned_data <- cleaned_data %>%
  mutate(
    Interaction_1 = Crossing * Traffic_Signal,
```

```r
    Interaction_2 = Traffic_Signal * Amenity,
    Interaction_3 = Crossing * Amenity,
    Interaction_4 = Amenity * Traffic_Signal * Crossing
  )


# SCALE NUMERICS
numeric_columns <- c("Temperature.F.", "Distance.mi.", "Wind_Speed.mph.", "Precipitation.in.")
cleaned_data[numeric_columns] <- lapply(cleaned_data[numeric_columns], as.numeric)
cleaned_data[numeric_columns] <- scale(cleaned_data[numeric_columns])
# CONVERT INT
categorical_columns <- c()
cleaned_data[categorical_columns] <- lapply(cleaned_data[categorical_columns], function(x) as.integer(fa

# CONVERT TO ORDERED FACTOR
cleaned_data$Severity <- factor(cleaned_data$Severity, levels = 1:4, ordered = TRUE)

# prepare matrices
X <- cleaned_data %>%
  select(-Severity) %>%
  mutate(across(everything(), as.numeric)) %>%
  as.matrix()
y <- as.integer(cleaned_data$Severity)

# BAYES ORDINAL MODEL
fit_ordinal <- brm(
  Severity ~ Crossing + Amenity + Traffic_Signal + Temperature.F. +
    Junction + Distance.mi. + Interaction_1 + Interaction_2 + Interaction_3 + Interaction_4 + Wind_Spee
  data = cleaned_data,
  family = cumulative(),
  chains = 4, iter = 2000, warmup = 1000,
  prior = c( # uninformative priors
    set_prior("normal(0, 1)", class = "b"),          # coefs
    set_prior("cauchy(0, 2)", class = "Intercept")  # thresholds
  )
)
```

```
## Compiling Stan program...

## Start sampling

##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0.000692 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 6.92 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
```

```
## Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 10.801 seconds (Warm-up)
## Chain 1:                7.826 seconds (Sampling)
## Chain 1:                18.627 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 0.000543 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 5.43 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 2:
## Chain 2:  Elapsed Time: 11.038 seconds (Warm-up)
## Chain 2:                7.467 seconds (Sampling)
## Chain 2:                18.505 seconds (Total)
## Chain 2:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 3).
## Chain 3:
## Chain 3: Gradient evaluation took 0.000704 seconds
## Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 7.04 seconds.
## Chain 3: Adjust your expectations accordingly!
## Chain 3:
## Chain 3:
## Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
```

```
## Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 3:
## Chain 3:  Elapsed Time: 9.026 seconds (Warm-up)
## Chain 3:                7.256 seconds (Sampling)
## Chain 3:                16.282 seconds (Total)
## Chain 3:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 4).
## Chain 4:
## Chain 4: Gradient evaluation took 0.000628 seconds
## Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 6.28 seconds.
## Chain 4: Adjust your expectations accordingly!
## Chain 4:
## Chain 4:
## Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 4:
## Chain 4:  Elapsed Time: 11.387 seconds (Warm-up)
## Chain 4:                8.4 seconds (Sampling)
## Chain 4:                19.787 seconds (Total)
## Chain 4:
```

```r
summary(fit_ordinal)
```

```
##  Family: cumulative
##   Links: mu = logit; disc = identity
## Formula: Severity ~ Crossing + Amenity + Traffic_Signal + Temperature.F. + Junction + Distance.mi. +
##     Data: cleaned_data (Number of observations: 1600)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Regression Coefficients:
##                 Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]       -1.43      0.07    -1.57    -1.30 1.00     4126     2805
## Intercept[2]       -0.17      0.06    -0.29    -0.05 1.00     5294     3821
## Intercept[3]        1.04      0.07     0.91     1.17 1.00     5527     3809
## Crossing           -0.94      0.26    -1.45    -0.41 1.00     3701     3247
## Amenity             0.32      0.56    -0.77     1.41 1.00     5386     3477
## Traffic_Signal     -0.94      0.17    -1.27    -0.61 1.00     4611     3008
## Temperature.F.     -0.38      0.05    -0.47    -0.29 1.00     5464     2899
## Junction            0.40      0.17     0.08     0.73 1.00     5343     3081
```

```
## Distance.mi.            0.37      0.06      0.25      0.50 1.00     5795      2921
## Interaction_1           0.36      0.34     -0.29      1.01 1.00     3585      2923
## Interaction_2          -0.99      0.73     -2.41      0.44 1.00     5070      3267
## Interaction_3           0.19      0.73     -1.20      1.63 1.00     5251      3108
## Interaction_4          -0.42      0.84     -2.09      1.20 1.00     5209      3004
## Wind_Speed.mph.        -0.01      0.05     -0.11      0.08 1.00     4975      2784
## Precipitation.in.       0.03      0.05     -0.06      0.12 1.00     5792      2937
##
## Further Distributional Parameters:
##      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## disc     1.00      0.00     1.00     1.00   NA       NA       NA
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```r
# predicted probabilities for each class (level of severity)
fitted_probs <- fitted(fit_ordinal, scale = "response")
str(fitted_probs)
```

```
##  num [1:1600, 1:4, 1:4] 0.471 0.633 0.309 0.643 0.284 ...
##  - attr(*, "dimnames")=List of 3
##   ..$ : NULL
##   ..$ : chr [1:4] "Estimate" "Est.Error" "Q2.5" "Q97.5"
##   ..$ : chr [1:4] "P(Y = 1)" "P(Y = 2)" "P(Y = 3)" "P(Y = 4)"
```

```r
# extract estimates of predicted probabilities
fitted_probs_estimate <- fitted_probs[, "Estimate", ]

# check structure again for indexing
str(fitted_probs_estimate)
```

```
##  num [1:1600, 1:4] 0.471 0.633 0.309 0.643 0.284 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:4] "P(Y = 1)" "P(Y = 2)" "P(Y = 3)" "P(Y = 4)"
```

```r
# make class prediction
y_pred_class <- apply(fitted_probs_estimate, 1, function(x) {
  which.max(x)  # index of highest prob class
})

# misclass rate
misclassification_rate <- mean(y != y_pred_class)
print(cat("Misclassification rate:", misclassification_rate, "\n"))
```

```
## Misclassification rate: 0.644375
## NULL
```

```r
# Extract posterior samples using as_draws
posterior_samples <- as_draws_df(fit_ordinal)
```

```r
# Ensure it's a data frame and extract coefficients (prefix 'b_')
beta_samples <- posterior_samples %>%
  select(starts_with("b_"))  # Extract coefficients (prefix 'b_')
```

Warning: Dropping 'draws_df' class as required metadata was removed.

```r
# Convert to matrix for plotting
beta_matrix <- as.matrix(beta_samples)

# Plot posterior distributions using mcmc_areas
mcmc_areas(beta_matrix) +
  ggtitle("Posterior Distributions of Beta Coefficients") +
  theme_minimal()
```


Posterior Distributions of Beta Coefficients
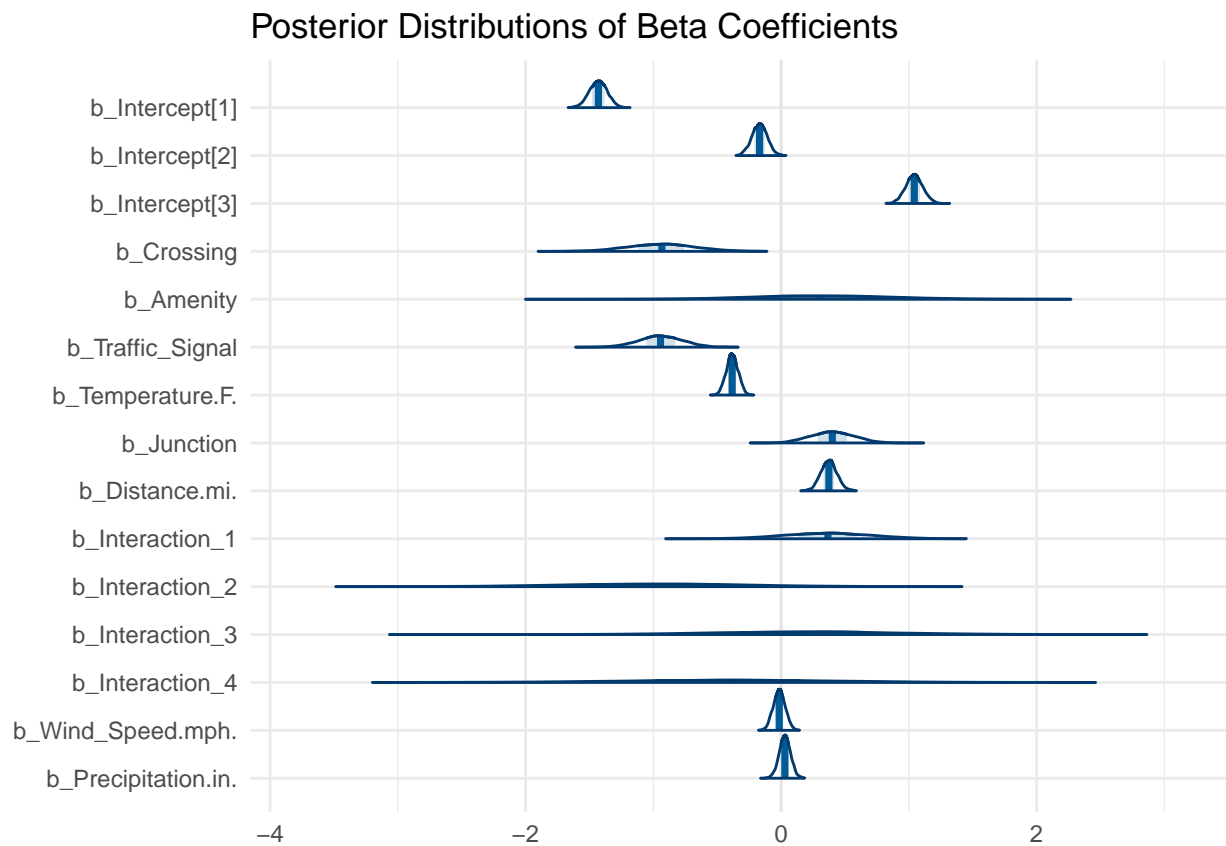
```r
# Summarize sampled data
summary(sampled_data)
```

```
##      Severity      Distance.mi.     Description          Start_Lat
##  Min.   :1.00   Min.   : 0.0000   Length:1600        Min.   :25.58
##  1st Qu.:1.75   1st Qu.: 0.0000   Class :character   1st Qu.:33.63
##  Median :2.50   Median : 0.0000   Mode  :character   Median :36.36
##  Mean   :2.50   Mean   : 0.6759                      Mean   :36.57
##  3rd Qu.:3.25   3rd Qu.: 0.4500                      3rd Qu.:40.26
##  Max.   :4.00   Max.   :31.2500                      Max.   :48.42
##    Start_Lng              City              County              State
```

```
##   Min.    :-124.16   Length:1600      Length:1600      Length:1600
##   1st Qu.:-107.73   Class :character   Class :character   Class :character
##   Median : -86.22   Mode  :character   Mode  :character   Mode  :character
##   Mean   : -92.76
##   3rd Qu.: -80.25
##   Max.   : -70.21
##    Start_Time          End_Time          Timezone         Temperature.F.
##   Length:1600       Length:1600       Length:1600       Min.   :-20.90
##   Class :character   Class :character   Class :character   1st Qu.: 51.00
##   Mode  :character   Mode  :character   Mode  :character   Median : 66.00
##                                                            Mean   : 63.13
##                                                            3rd Qu.: 77.00
##                                                            Max.   :109.00
##   Wind_Chill.F.      Humidity...       Pressure.in.     Visibility.mi.
##   Min.   :-48.50    Min.   :  5.00    Min.   :20.56    Min.   : 0.000
##   1st Qu.: 51.00    1st Qu.: 49.00    1st Qu.:29.10    1st Qu.:10.000
##   Median : 66.00    Median : 68.00    Median :29.58    Median :10.000
##   Mean   : 61.81    Mean   : 65.21    Mean   :29.23    Mean   : 9.083
##   3rd Qu.: 77.00    3rd Qu.: 84.00    3rd Qu.:29.91    3rd Qu.:10.000
##   Max.   :109.00    Max.   :100.00    Max.   :30.62    Max.   :20.000
##   Wind_Direction    Wind_Speed.mph.  Precipitation.in.  Weather_Condition
##   Length:1600       Min.   : 0.000    Min.   :0.000000    Length:1600
##   Class :character   1st Qu.: 3.000    1st Qu.:0.000000    Class :character
##   Mode  :character   Median : 7.000    Median :0.000000    Mode  :character
##                     Mean   : 7.435    Mean   :0.006044
##                     3rd Qu.:10.000    3rd Qu.:0.000000
##                     Max.   :36.000    Max.   :0.460000
##     Amenity            Bump             Crossing          Junction
##   Length:1600       Length:1600       Length:1600       Length:1600
##   Class :character   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     No_Exit        Traffic_Calming   Traffic_Signal       Duration
##   Length:1600       Length:1600       Length:1600       Min.   :    7.067
##   Class :character   Class :character   Class :character   1st Qu.:   34.967
##   Mode  :character   Mode  :character   Mode  :character   Median :   62.600
##                                                            Mean   :  107.281
##                                                            3rd Qu.:  109.554
##                                                            Max.   :10275.500
##   Time_of_Day
##   Length:1600
##   Class :character
##   Mode  :character
##
##
##
```

```r
# Quantitative variables in sampled data
sampled_data %>%
  select(where(is.numeric)) %>%
  summary()
```

```
##      Severity        Distance.mi.        Start_Lat         Start_Lng
##  Min.   :1.00    Min.   : 0.0000    Min.   :25.58    Min.   :-124.16
##  1st Qu.:1.75    1st Qu.: 0.0000    1st Qu.:33.63    1st Qu.:-107.73
##  Median :2.50    Median : 0.0000    Median :36.36    Median : -86.22
##  Mean   :2.50    Mean   : 0.6759    Mean   :36.57    Mean   : -92.76
##  3rd Qu.:3.25    3rd Qu.: 0.4500    3rd Qu.:40.26    3rd Qu.: -80.25
##  Max.   :4.00    Max.   :31.2500    Max.   :48.42    Max.   : -70.21
##  Temperature.F.    Wind_Chill.F.      Humidity...       Pressure.in.
##  Min.   :-20.90    Min.   :-48.50    Min.   :  5.00    Min.   :20.56
##  1st Qu.: 51.00    1st Qu.: 51.00    1st Qu.: 49.00    1st Qu.:29.10
##  Median : 66.00    Median : 66.00    Median : 68.00    Median :29.58
##  Mean   : 63.13    Mean   : 61.81    Mean   : 65.21    Mean   :29.23
##  3rd Qu.: 77.00    3rd Qu.: 77.00    3rd Qu.: 84.00    3rd Qu.:29.91
##  Max.   :109.00    Max.   :109.00    Max.   :100.00    Max.   :30.62
##  Visibility.mi.    Wind_Speed.mph.   Precipitation.in.    Duration
##  Min.   : 0.000    Min.   : 0.000    Min.   :0.000000    Min.   :     7.067
##  1st Qu.:10.000    1st Qu.: 3.000    1st Qu.:0.000000    1st Qu.:    34.967
##  Median :10.000    Median : 7.000    Median :0.000000    Median :    62.600
##  Mean   : 9.083    Mean   : 7.435    Mean   :0.006044    Mean   :   107.281
##  3rd Qu.:10.000    3rd Qu.:10.000    3rd Qu.:0.000000    3rd Qu.:   109.554
##  Max.   :20.000    Max.   :36.000    Max.   :0.460000    Max.   :10275.500
```

```r
# Categorical variables in sampled data
sampled_data %>%
  select(where(is.factor)) %>%
  lapply(table)
```

```
## named list()
```

```r
# Accident Severity Distribution
ggplot(sampled_data, aes(x = factor(Severity))) +
  geom_bar(fill = "lightblue") +
  labs(title = "Accident Severity Distribution", x = "Severity", y = "Count")
```

Accident Severity Distribution
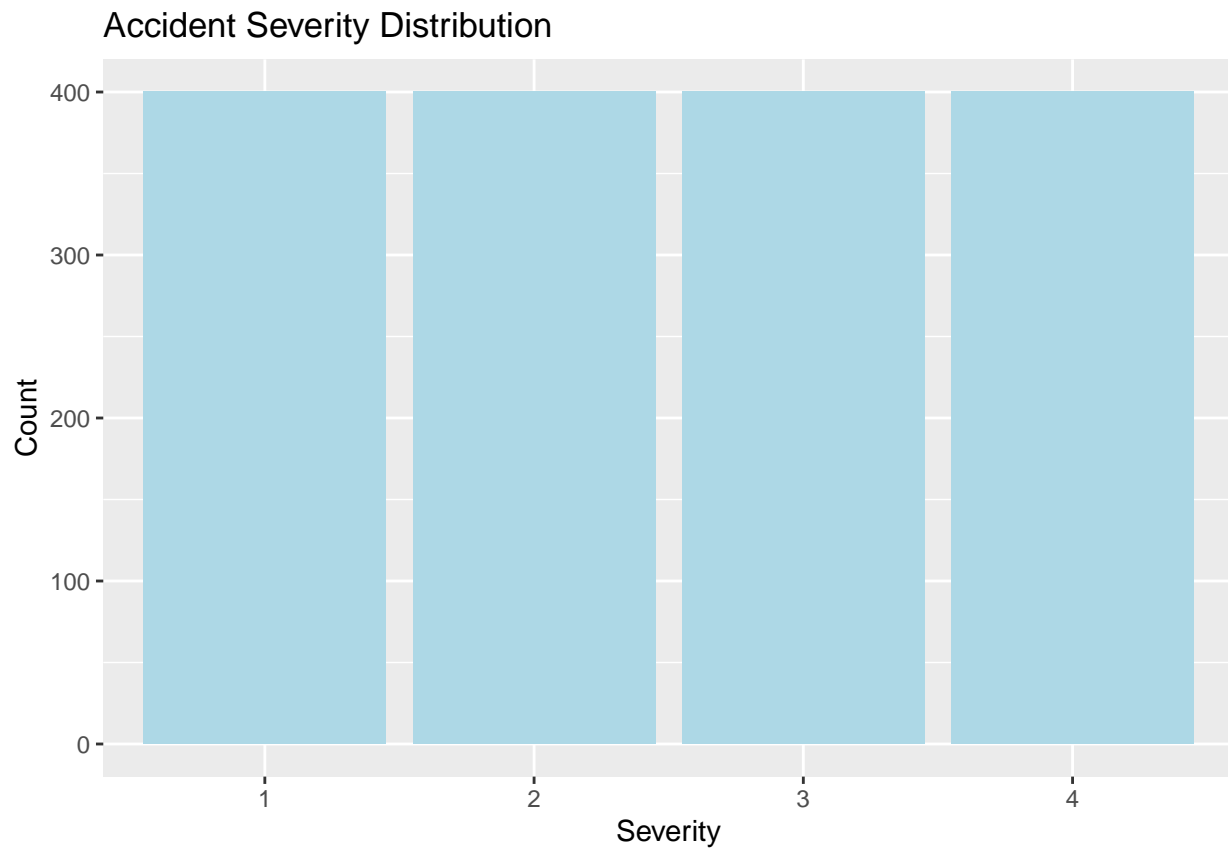

```r
# Accidents by Time of Day
ggplot(sampled_data, aes(x = Time_of_Day)) +
  geom_bar(fill = "orange") +
  labs(title = "Accidents by Time of Day", x = "Time of Day", y = "Count")
```

Accidents by Time of Day
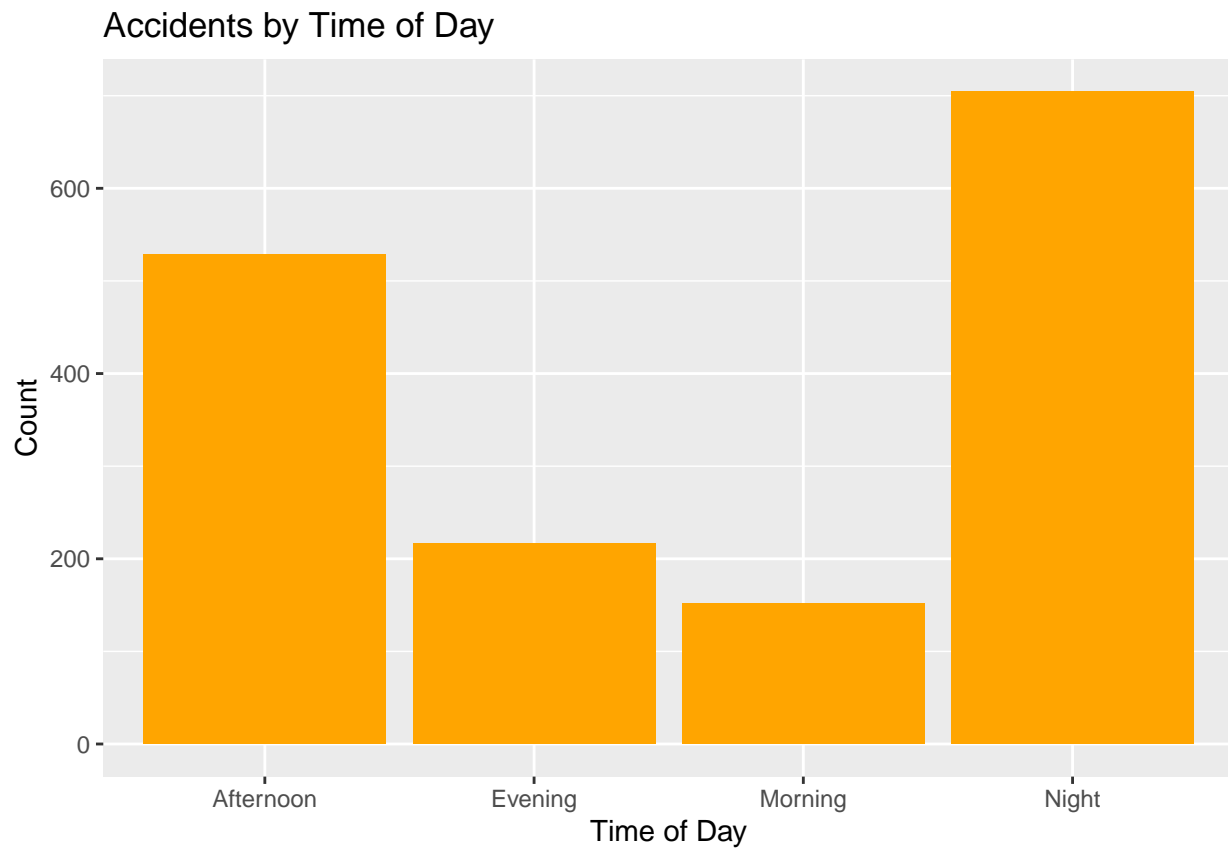

```r
# Accident Distance by Severity
ggplot(sampled_data, aes(x = factor(Severity), y = Distance.mi.)) +
  geom_boxplot(fill = "blue") +
  labs(title = "Accident Distance by Severity", x = "Severity", y = "Distance (miles)")
```

Accident Distance by Severity
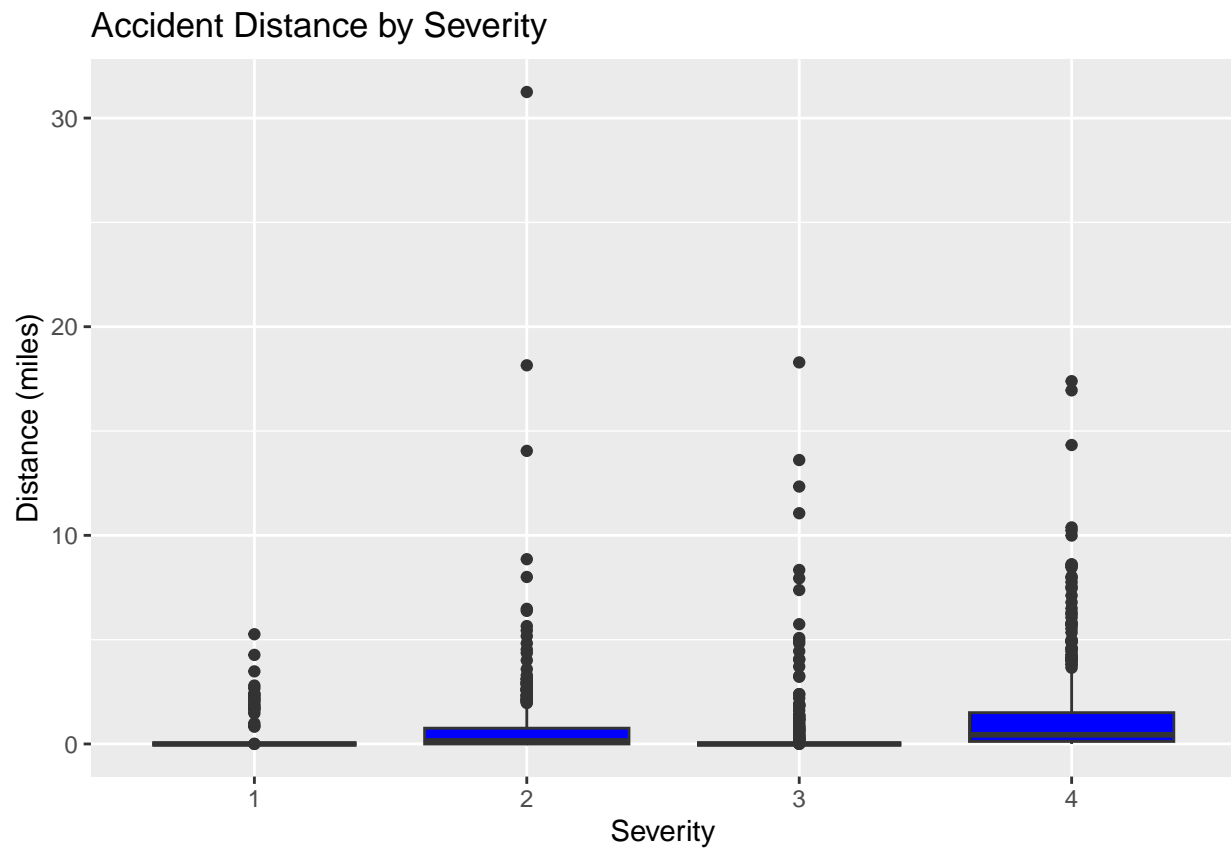

```r
# Accident Severity by Time of Day
ggplot(sampled_data, aes(x = Time_of_Day, fill = factor(Severity))) +
  geom_bar(position = "dodge") +
  labs(title = "Accident Severity by Time of Day", x = "Time of Day", y = "Count", fill = "Severity")
```

Accident Severity by Time of Day
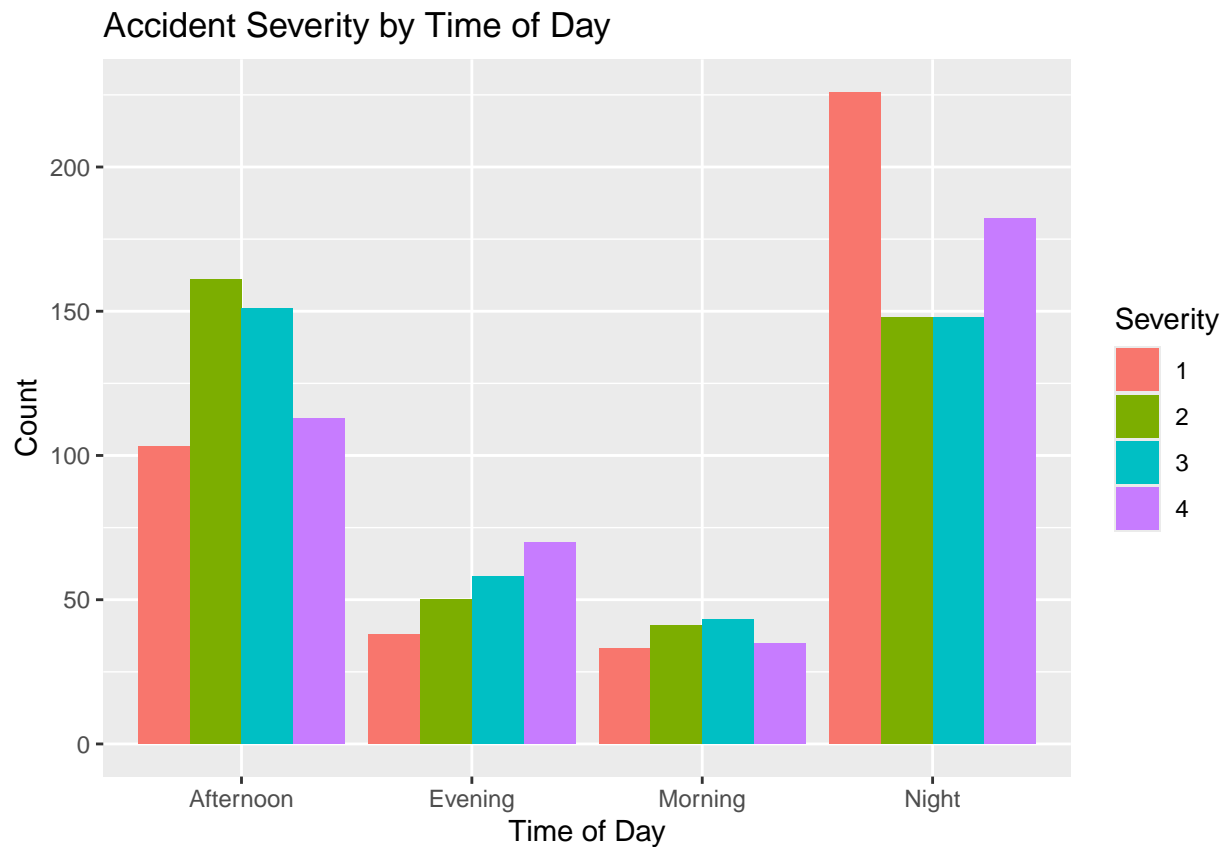
```
# Severity by Weather Condition
ggplot(sampled_data, aes(x = Weather_Condition, fill = factor(Severity))) +  theme_minimal() +
  theme(axis.text.y = element_text(size = 5))+  # Set smaller size for y-axis labels
  geom_bar(position = "dodge", width =2) + coord_flip()
```

## Warning: 'position_dodge()' requires non-overlapping x intervals.
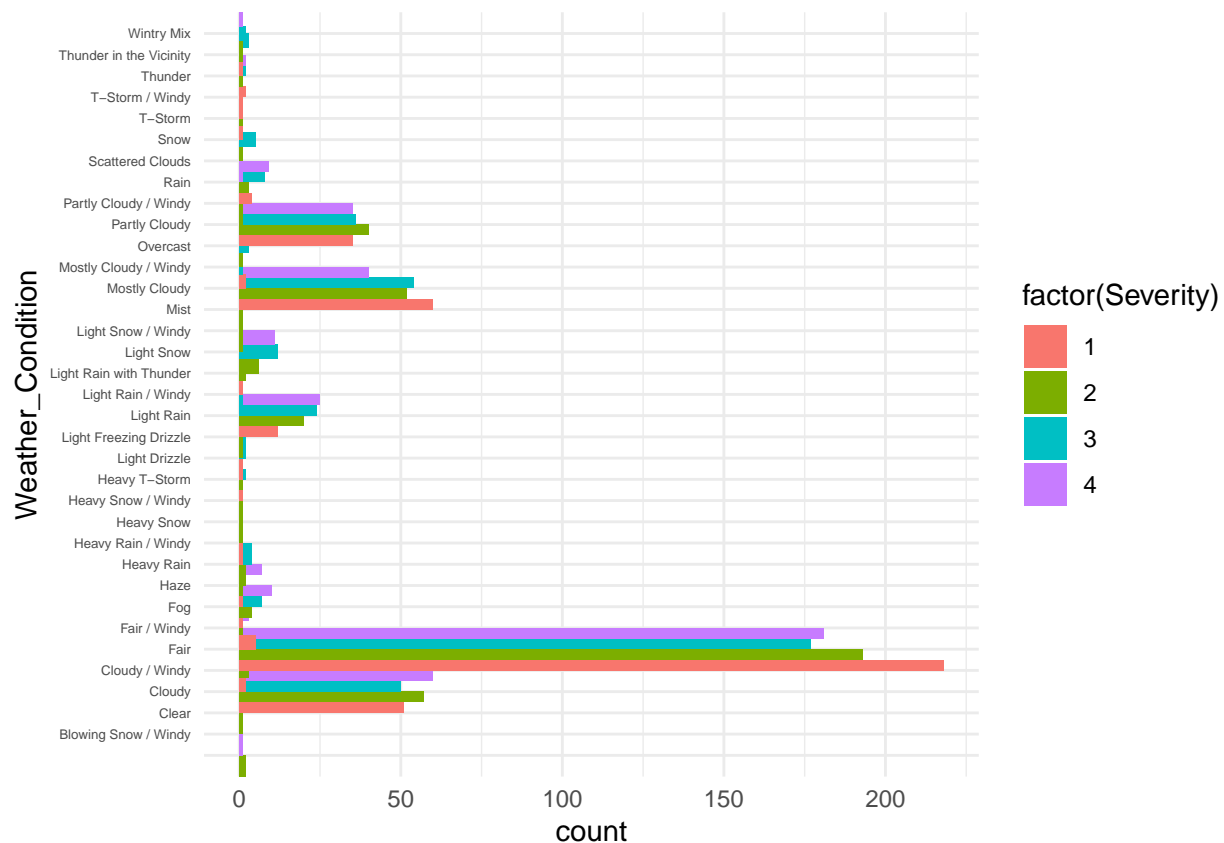
```
labs(title = "Severity by Weather Condition", x = "Weather Condition", y = "Count", fill = "Severity")
```

```
## $x
## [1] "Weather Condition"
##
## $y
## [1] "Count"
##
## $fill
## [1] "Severity"
##
## $title
## [1] "Severity by Weather Condition"
##
## attr(,"class")
## [1] "labels"
```

```
  # Severity by Weather Condition with smaller y-axis text
ggplot(sampled_data, aes(x = Weather_Condition, fill = factor(Severity))) +
  geom_bar(position = "dodge", width = 3) +  # Adjust bar width
  coord_flip() +
  labs(
    title = "Severity by Weather Condition",
    x = "Weather Condition",
    y = "Count",
    fill = "Severity"
```

```
  ) +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 5),   # Set smaller size for y-axis labels
    plot.title = element_text(hjust = 1, size = 10), # Center and enlarge title
    legend.position = "bottom" # Move legend to bottom
  )
```

Warning: 'position_dodge()' requires non-overlapping x intervals.
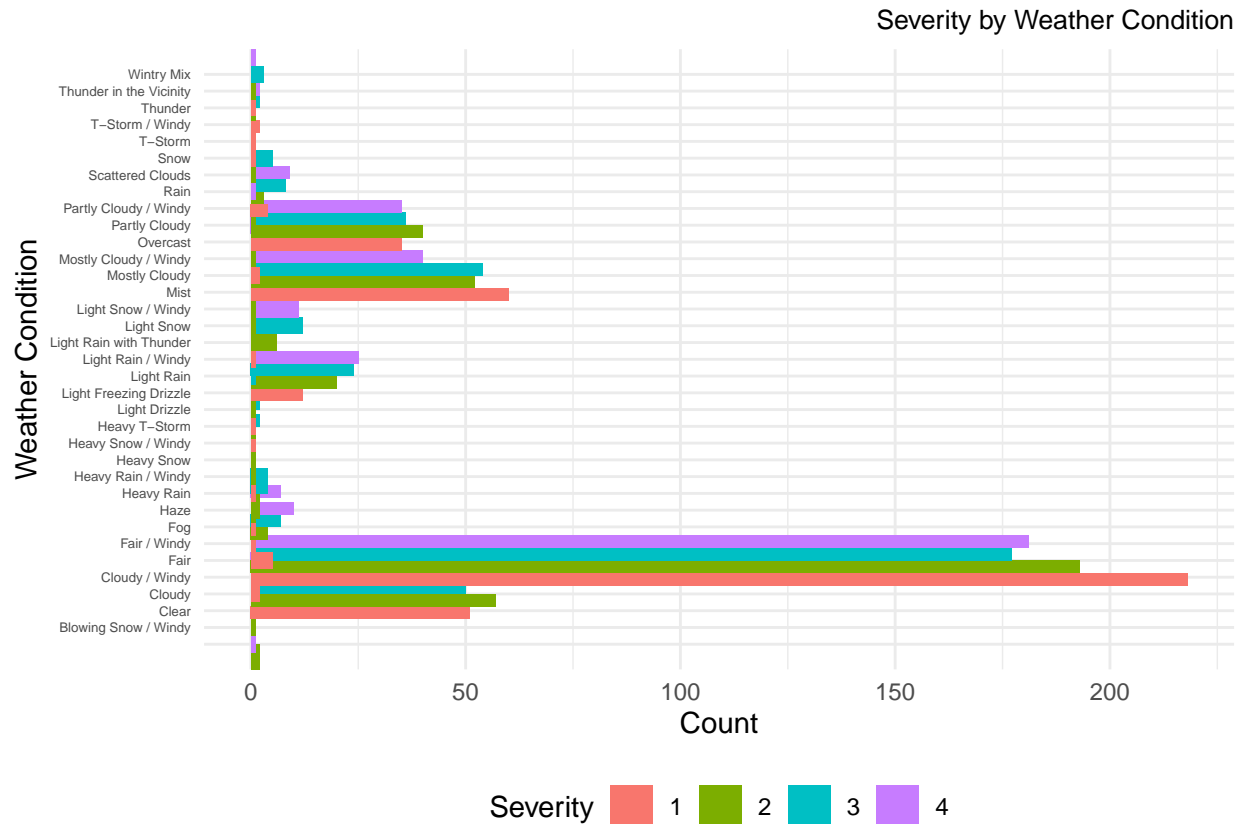

```
library(dplyr)


cleaned_data <- read.csv("sampled_data4.csv")

set.seed(123)
cleaned_data <- cleaned_data %>%
  select(Temperature.F., Wind_Chill.F., Humidity..., Pressure.in., Visibility.mi.,
         Wind_Direction, Wind_Speed.mph., Precipitation.in., Weather_Condition,
         Amenity, Bump, Crossing, Junction, No_Exit, Traffic_Calming, Traffic_Signal,
         Duration, Time_of_Day, Distance.mi.)



binary_columns <- c("Amenity", "Bump", "No_Exit", "Traffic_Calming", "Traffic_Signal", "Junction", "Cros
cleaned_data[binary_columns] <- lapply(cleaned_data[binary_columns], function(x)
  as.integer(factor(x, levels = c("False", "True"), labels = c(0, 1))))
```

```r
numeric_columns <- c("Temperature.F.", "Wind_Chill.F.", "Humidity...", "Wind_Speed.mph.",
                     "Distance.mi.", "Precipitation.in.", "Pressure.in.", "Duration")
cleaned_data[numeric_columns] <- lapply(cleaned_data[numeric_columns], as.numeric)
cleaned_data[numeric_columns] <- scale(cleaned_data[numeric_columns])


categorical_columns <- c("Weather_Condition", "Time_of_Day", "Wind_Direction","Visibility.mi.")
cleaned_data[categorical_columns] <- lapply(cleaned_data[categorical_columns], function(x) {as.integer(


cor_matrix <- cor(cleaned_data)
```

```
## Warning in cor(cleaned_data): the standard deviation is zero
```

```r
print(cor_matrix)
```

```
##                   Temperature.F. Wind_Chill.F. Humidity... Pressure.in.
## Temperature.F.        1.00000000   0.993699343 -0.32455817  0.094282144
## Wind_Chill.F.         0.99369934   1.000000000 -0.30333559  0.101627885
## Humidity...          -0.32455817  -0.303335587  1.00000000  0.257705440
## Pressure.in.          0.09428214   0.101627885  0.25770544  1.000000000
## Visibility.mi.        0.29368343   0.304761637 -0.34952869 -0.040650311
## Wind_Direction        0.10415997   0.075634200 -0.23687690 -0.044385003
## Wind_Speed.mph.      -0.03200121  -0.086636201 -0.19979591 -0.037287147
## Precipitation.in.    -0.04768744  -0.042464333  0.20141942  0.034721559
## Weather_Condition     0.01605342   0.002768528  0.09881530  0.028244761
## Amenity               0.01715374   0.019806373  0.01857040 -0.006382651
## Bump                          NA            NA          NA           NA
## Crossing              0.15113164   0.144835003 -0.14846322 -0.052968495
## Junction             -0.10343819  -0.103232547  0.06576664  0.070811223
## No_Exit               0.01016657   0.004350677 -0.03892398 -0.005120933
## Traffic_Calming       0.01531157   0.015731619  0.01566988  0.021392579
## Traffic_Signal        0.12541494   0.121342251 -0.13422796 -0.028271994
## Duration             -0.03659869  -0.035622972 -0.03224784 -0.090212235
## Time_of_Day          -0.22452515  -0.201785980  0.44485271  0.037351847
## Distance.mi.         -0.08692829  -0.092994646  0.02523017 -0.108215902
##                   Visibility.mi. Wind_Direction Wind_Speed.mph.
## Temperature.F.        0.293683430     0.10415997     -0.032001209
## Wind_Chill.F.         0.304761637     0.07563420     -0.086636201
## Humidity...          -0.349528694    -0.23687690     -0.199795907
## Pressure.in.         -0.040650311    -0.04438500     -0.037287147
## Visibility.mi.        1.000000000     0.05157625     -0.066412709
## Wind_Direction        0.051576248     1.00000000      0.439273589
## Wind_Speed.mph.      -0.066412709     0.43927359      1.000000000
## Precipitation.in.    -0.269923787    -0.02840456      0.029087806
## Weather_Condition    -0.154398133     0.04793765      0.134569149
## Amenity               0.008644525     0.02335392      0.001067824
## Bump                           NA             NA               NA
## Crossing              0.086798943     0.03583352      0.057572850
## Junction             -0.002893578    -0.03955994     -0.012469309
## No_Exit               0.015456219     0.04587538      0.041223777
```

```
## Traffic_Calming        0.007145648    -0.03424185    -0.031744373
## Traffic_Signal         0.089465757     0.04804786     0.047494595
## Duration              -0.001260607    -0.01402569    -0.016255514
## Time_of_Day           -0.054202286    -0.21339135    -0.294803096
## Distance.mi.          -0.040011993    -0.00102320     0.011422267
##                      Precipitation.in. Weather_Condition      Amenity Bump
## Temperature.F.         -0.047687440        0.0160534195  0.0171537422   NA
## Wind_Chill.F.          -0.042464333        0.0027685279  0.0198063733   NA
## Humidity...             0.201419418        0.0988153026  0.0185704027   NA
## Pressure.in.            0.034721559        0.0282447615 -0.0063826510   NA
## Visibility.mi.         -0.269923787       -0.1543981332  0.0086445252   NA
## Wind_Direction         -0.028404563        0.0479376480  0.0233539218   NA
## Wind_Speed.mph.         0.029087806        0.1345691486  0.0010678236   NA
## Precipitation.in.       1.000000000        0.1924731181  0.0179761092   NA
## Weather_Condition       0.192473118        1.0000000000  0.0005842388   NA
## Amenity                 0.017976109        0.0005842388  1.0000000000   NA
## Bump                            NA                  NA            NA    1
## Crossing               -0.016171676        0.0106326397  0.1189129639   NA
## Junction               -0.029692387       -0.0394022218 -0.0306324207   NA
## No_Exit                -0.008034638        0.0280391449 -0.0043560271   NA
## Traffic_Calming        -0.006558201       -0.0231524422  0.3520012516   NA
## Traffic_Signal         -0.055107119       -0.0347201473  0.0859646234   NA
## Duration               -0.015851379       -0.0403004779 -0.0115113124   NA
## Time_of_Day             0.019499364       -0.0537836854 -0.0109569469   NA
## Distance.mi.            0.007106926        0.0273050084 -0.0140300736   NA
##                          Crossing       Junction       No_Exit Traffic_Calming
## Temperature.F.         0.151131640   -0.103438195   0.010166570      0.015311575
## Wind_Chill.F.          0.144835003   -0.103232547   0.004350677      0.015731619
## Humidity...           -0.148463223    0.065766635  -0.038923981      0.015669876
## Pressure.in.          -0.052968495    0.070811223  -0.005120933      0.021392579
## Visibility.mi.         0.086798943   -0.002893578   0.015456219      0.007145648
## Wind_Direction         0.035833519   -0.039559939   0.045875382     -0.034241848
## Wind_Speed.mph.        0.057572850   -0.012469309   0.041223777     -0.031744373
## Precipitation.in.     -0.016171676   -0.029692387  -0.008034638     -0.006558201
## Weather_Condition      0.010632640   -0.039402222   0.028039145     -0.023152442
## Amenity                0.118912964   -0.030632421  -0.004356027      0.352001252
## Bump                           NA             NA            NA               NA
## Crossing               1.000000000   -0.104605548   0.073661976      0.096662765
## Junction              -0.104605548    1.000000000   0.038594301     -0.010782650
## No_Exit                0.073661976    0.038594301   1.000000000     -0.001533327
## Traffic_Calming        0.096662765   -0.010782650  -0.001533327      1.000000000
## Traffic_Signal         0.560216720   -0.122681010   0.056083219     -0.016293660
## Duration              -0.005295649   -0.006422104  -0.003757581     -0.006918424
## Time_of_Day            0.036536921    0.032486744  -0.010156350      0.009442655
## Distance.mi.          -0.105272584    0.035092632  -0.014097181     -0.012607210
##                      Traffic_Signal     Duration   Time_of_Day Distance.mi.
## Temperature.F.           0.12541494  -0.036598685  -0.224525146  -0.086928287
## Wind_Chill.F.            0.12134225  -0.035622972  -0.201785980  -0.092994646
## Humidity...             -0.13422796  -0.032247839   0.444852706   0.025230172
## Pressure.in.            -0.02827199  -0.090212235   0.037351847  -0.108215902
## Visibility.mi.           0.08946576  -0.001260607  -0.054202286  -0.040011993
## Wind_Direction           0.04804786  -0.014025692  -0.213391351  -0.001023200
## Wind_Speed.mph.          0.04749460  -0.016255514  -0.294803096   0.011422267
## Precipitation.in.       -0.05510712  -0.015851379   0.019499364   0.007106926
```

```
## Weather_Condition      -0.03472015 -0.040300478 -0.053783685  0.027305008
## Amenity                 0.08596462 -0.011511312 -0.010956947 -0.014030074
## Bump                            NA           NA           NA           NA
## Crossing                0.56021672 -0.005295649  0.036536921 -0.105272584
## Junction               -0.12268101 -0.006422104  0.032486744  0.035092632
## No_Exit                 0.05608322 -0.003757581 -0.010156350 -0.014097181
## Traffic_Calming        -0.01629366 -0.006918424  0.009442655 -0.012607210
## Traffic_Signal          1.00000000 -0.057906140  0.014098666 -0.133067695
## Duration               -0.05790614  1.000000000  0.017030315  0.050348048
## Time_of_Day             0.01409867  0.017030315  1.000000000  0.016222202
## Distance.mi.           -0.13306769  0.050348048  0.016222202  1.000000000
```

```r
heatmap(cor_matrix,
        main = "Correlation Matrix",
        col = colorRampPalette(c("blue", "white", "red"))(100),
        scale = "none",
        margins = c(10, 10),
        Rowv = NA,    # readability
        Colv = NA,
)
```

Correlation Matrix
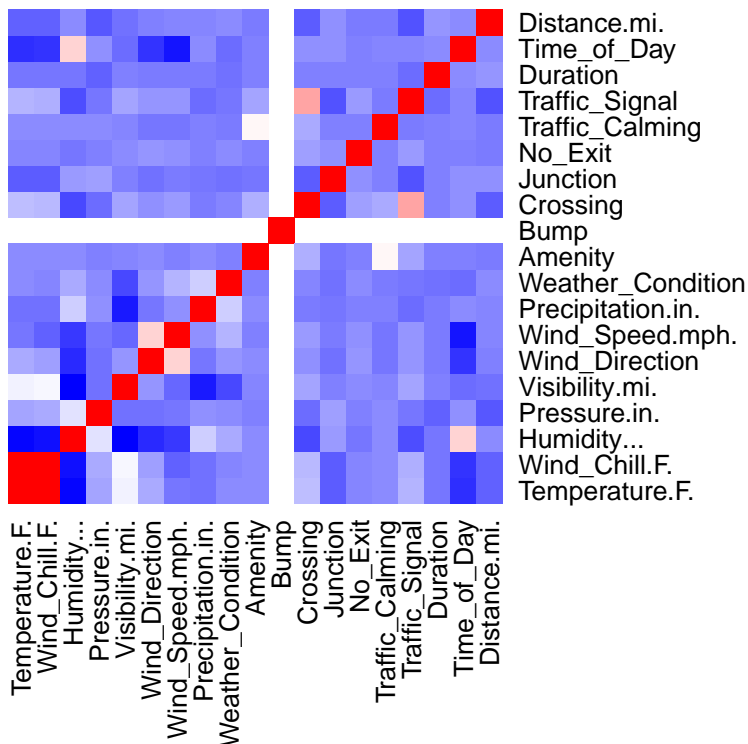
MCMC tracing, posterior and Posterior predictive check

```r
library(bayesplot)


# Use mcmc_trace with brms model (fit_ordinal)
# Extract draws as an array
trace_data <- as_draws_array(fit_ordinal)
```

```r
# Plot trace for specific parameters
mcmc_trace(trace_data, pars = c("b_Crossing", "b_Amenity", "b_Traffic_Signal","b_Temperature.F.", "b_Ju
  ggtitle("Trace Plots for Selected Parameters") +
  theme_minimal()
```

Trace Plots for Selected Parameters
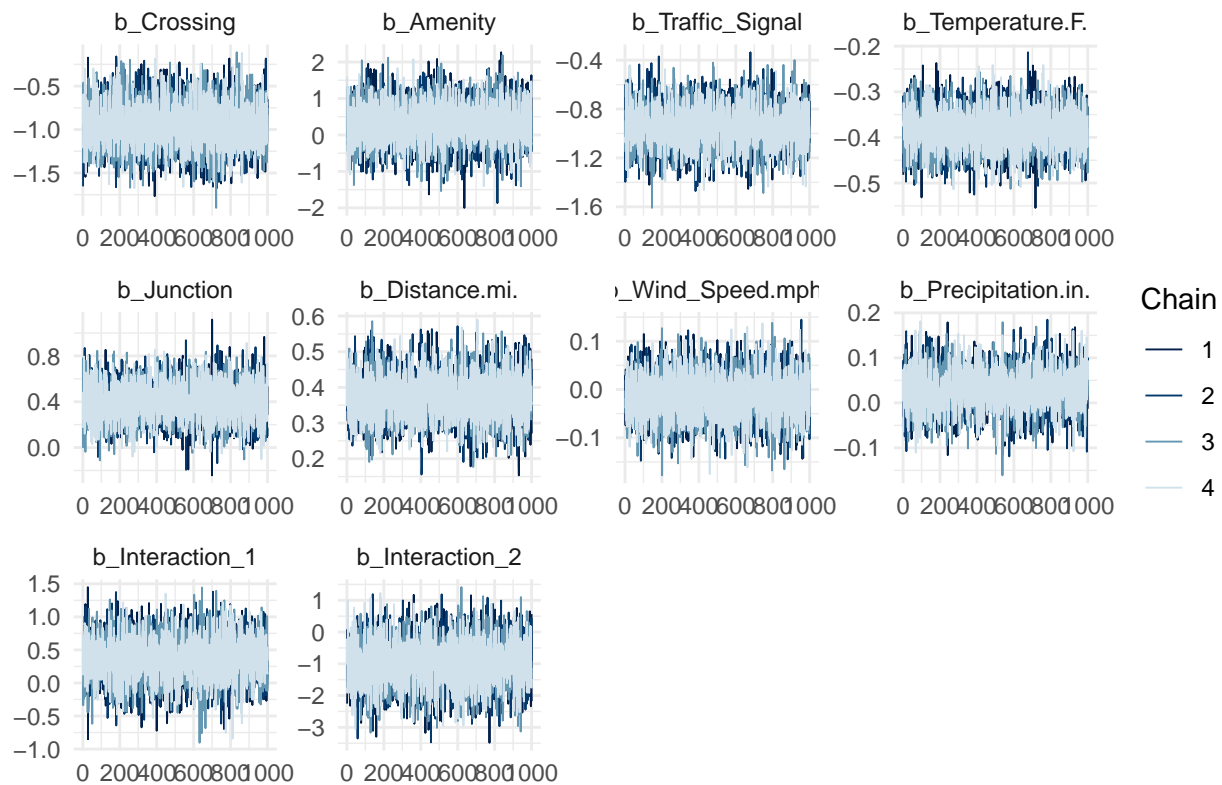

```r
# Posterior distributions for coefficients
posterior_samples <- as_draws_df(fit_ordinal)

mcmc_areas(
  posterior_samples %>% select(starts_with("b_")),
  prob = 0.9  # 90% credible intervals
) +
  ggtitle("Posterior Distributions with 90% Credible Intervals") +
  theme_minimal()
```
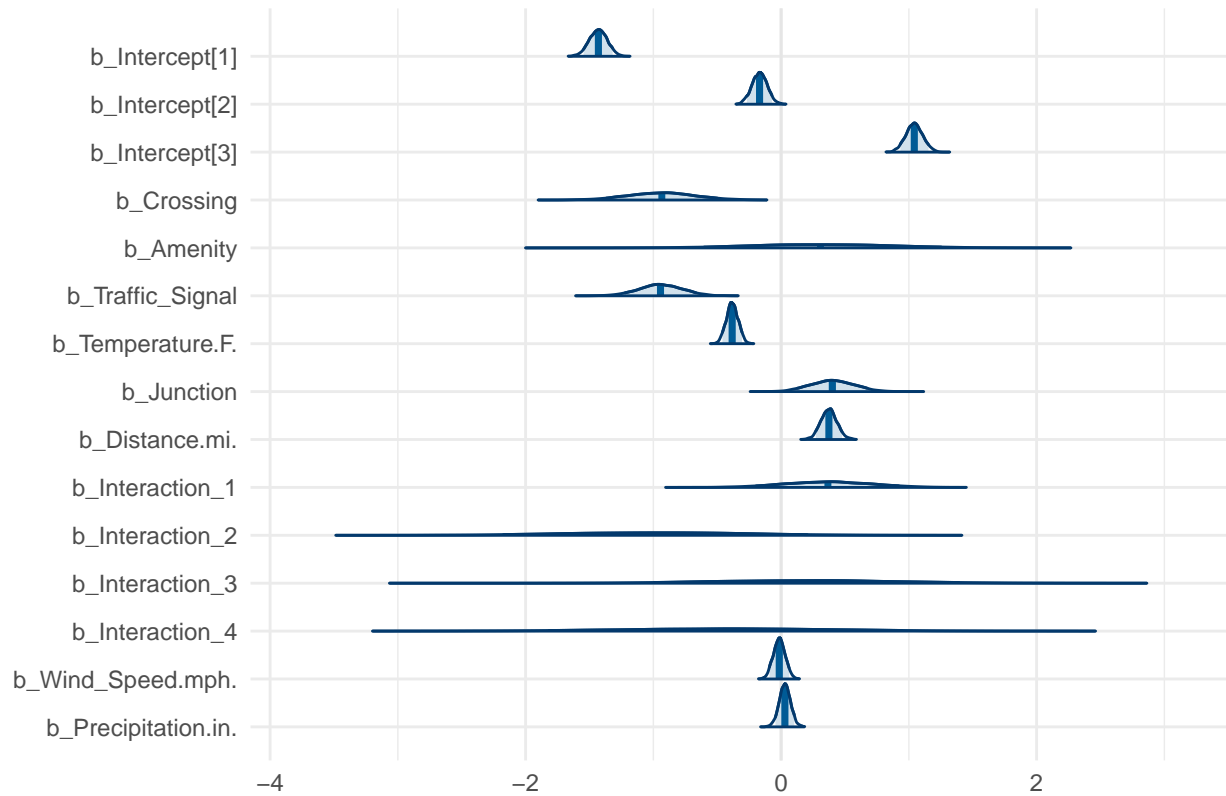
```
## Warning: Dropping 'draws_df' class as required metadata was removed.
```
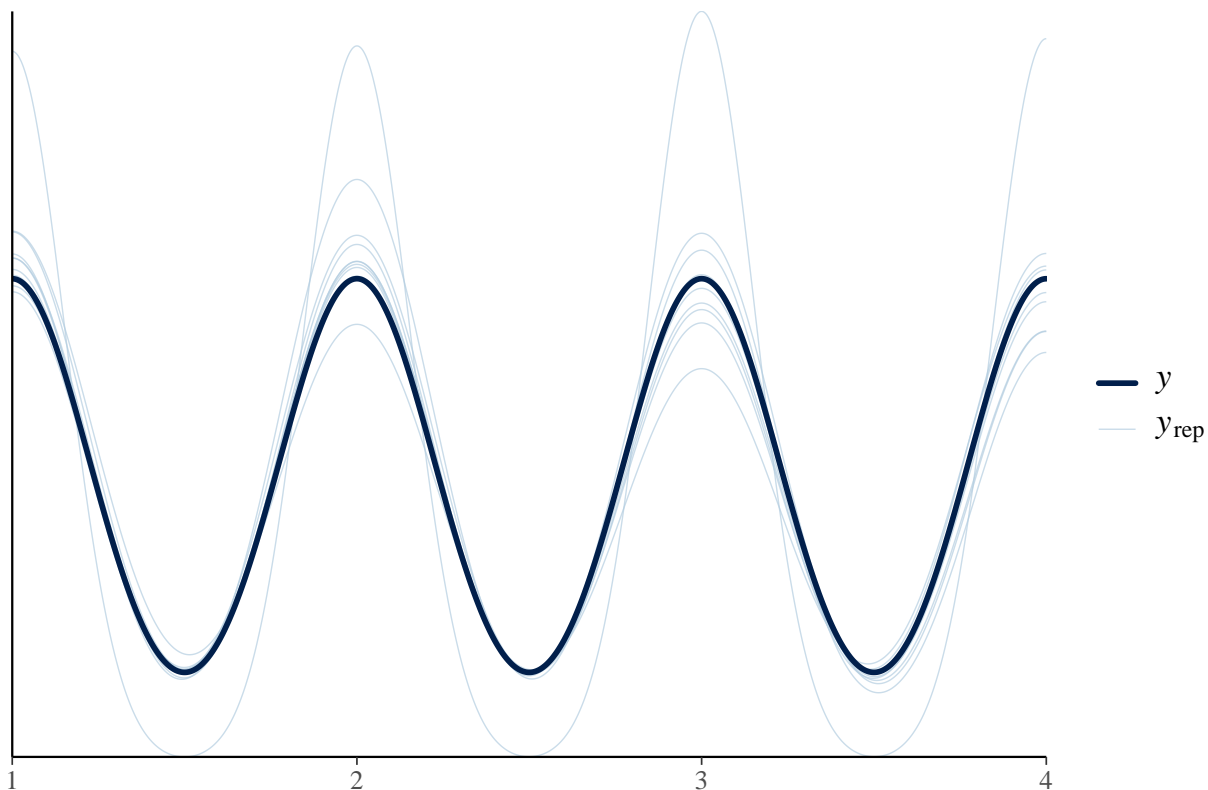
Posterior Distributions with 90% Credible Intervals


```r
# Posterior predictive check
pp_check(fit_ordinal) +
  ggtitle("Posterior Predictive Check")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

Posterior Predictive Check


```r
# Rhat and ESS diagnostics
fit_summary <- summary(fit_ordinal)

# Check fixed effects (coefficients)
fit_summary$fixed
```

```
##                     Estimate   Est.Error      l-95% CI      u-95% CI        Rhat
## Intercept[1]      -1.43134389  0.06976052  -1.56919791  -1.29736901  1.0007622
## Intercept[2]      -0.16846802  0.05906931  -0.28574553  -0.05325487  1.0003369
## Intercept[3]       1.04263690  0.06644558   0.91151714   1.17464078  0.9996648
## Crossing          -0.93668079  0.26331365  -1.44861540  -0.41397810  1.0013098
## Amenity            0.31833520  0.56481455  -0.76702779   1.41084989  1.0004072
## Traffic_Signal    -0.94185723  0.16858440  -1.27378888  -0.61185019  1.0019439
## Temperature.F.    -0.38254387  0.04605296  -0.47185933  -0.29280917  1.0018383
## Junction           0.40093573  0.16798541   0.08180619   0.72908065  1.0029220
## Distance.mi.       0.37385018  0.06283902   0.25392514   0.49798427  1.0007285
## Interaction_1      0.36355243  0.33925726  -0.29144428   1.01217938  1.0002246
## Interaction_2     -0.99313884  0.72971799  -2.41091659   0.44023687  1.0017688
## Interaction_3      0.18995190  0.72746936  -1.19712131   1.63144167  1.0019447
## Interaction_4     -0.41858021  0.84208297  -2.09434698   1.19884706  1.0007639
## Wind_Speed.mph.   -0.01440143  0.04686230  -0.10515087   0.07792519  0.9998719
## Precipitation.in.  0.02982745  0.04525230  -0.06136868   0.11764962  0.9997546
##                    Bulk_ESS Tail_ESS
## Intercept[1]       4126.440 2804.810
## Intercept[2]       5293.996 3820.974
## Intercept[3]       5526.687 3809.139
## Crossing           3700.906 3247.270
```

```
## Amenity          5386.158 3477.425
## Traffic_Signal   4610.565 3007.941
## Temperature.F.   5463.845 2898.954
## Junction         5343.124 3081.091
## Distance.mi.     5795.454 2921.169
## Interaction_1    3585.133 2922.859
## Interaction_2    5070.387 3266.876
## Interaction_3    5250.958 3108.117
## Interaction_4    5209.206 3004.240
## Wind_Speed.mph.  4975.110 2783.561
## Precipitation.in. 5792.402 2937.486
```

```r
# Check Rhat values (should be close to 1 for convergence)
fit_summary$rhat
```

```
## NULL
```