Sunsetter Seasonal Analysis

Skyler Hauser

2023-06-27

```
install.packages(c("DBI", "odbc", "dplyr"))
```

```
## Installing packages into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
install.packages("RPostgres")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)

library(dplyr)
library(knitr)
library(dplyr)

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(DBI)
library(odbc)
library(dplyr)
library(RPostgres)
```

```
con <- dbCanConnect(RPostgres::Postgres(),
                    dbname = "rfi",
```

```
                    port = 5432,
                    user = "nycfmexp",
                    password = "Zxcv1234",
                    host="rfi-production-cluster.cluster-cqx4czau66cg.us-east-1.rds.amazonaws.com")
con
```

```
## [1] TRUE
```

```
con_1 <- dbConnect(RPostgres::Postgres(),
                   dbname = "rfi",
                    port = 5432,
                    user = "nycfmexp",
                    password = "Zxcv1234",
                    host="rfi-production-cluster.cluster-cqx4czau66cg.us-east-1.rds.amazonaws.com")
con_1
```

```
## <PqConnection> rfi@rfi-production-cluster.cluster-cqx4czau66cg.us-east-1.rds.amazonaws.com:5432
```

```
dbListTables(con_1)
```

```
##  [1] "account_web_projects"        "account_api_keys"
##  [3] "additional_accounts"         "advertisers"
##  [5] "ar_internal_metadata"        "campaign_metrics_settings"
##  [7] "accounts"                    "core_buys"
##  [9] "core_spots"                  "cleared_spots"
## [11] "campaigns"                   "dayparts"
## [13] "flights"                     "insertion_orders"
## [15] "national_universe_estimates" "nielsen_c3_data"
## [17] "nielsen_daily_l3"            "nielsen_daily_overnight"
## [19] "nielsen_l3_data"             "nielsen_on_data"
## [21] "core_styles"                 "core_style_ratings"
## [23] "products"                    "stations"
## [25] "rfis"                        "schema_migrations"
## [27] "rfi_stations"                "rate_cards"
## [29] "station_dayparts"            "nielsens"
## [31] "order_items"                 "ratings"
## [33] "web_visit_results"           "spots"
## [35] "web_event_results"           "user_passwords"
## [37] "user_authorizations"         "user_roles"
## [39] "tokens"                      "users"
```

```
account_web_projects <- dbGetQuery(con_1, "SELECT * FROM account_web_projects")
account_web_projects
```

```
##     id account_id project_id                     name
## 1    1         39   16261128              Grayscale_01
## 2    2        137  133197832      StellaAndChewys_01\t
## 3    3        169   23619592               FreshPet_01
## 4    4        170   85985288 AgelessMale_NewVitality01
## 5    5        170  122695688      SuperBetaProstrate01
## 6    6        171   65562632               BalmLabs_01
## 7    7        172   53372928               Centipede01
## 8   10        174   31488008                Joyburst01
## 9   11        175   34609160              KidneyCOP_01
## 10  12        138   35151880               KumAndGo_01
## 11  13        176   46155784                    LUS_01
```

```
## 12  14        177   95854592                  Petro01
## 13  15        178   91258888       RejuvenateMuscle_01
## 14  16        136  112492544                slimfast01
## 15  17        179   53999616                 Sunsetter
## 16  18        180   27807752              WaterGuru_01
## 17   8        181  127420424             Speedhorse_01
## 18  34        203   78688264               OptiNail_01
## 19  35        203   79212552              FungiNail_01
## 20  36        204   44083208            HumidyFlame_01
## 21  67        204   44607496 Hy-ImpactMassagePillow_01
## 22  68        173  121968640              Hydroxycut02
## 23  69        235  130070536          PacificSource_01
## 24 100        268   13680648              Crunchcup_01
## 25 133        301   82884616               Opositiv_01
## 26 134        301   82360328            FloVitamins_01
##                             api_key          created_at          updated_at
## 1  LS4ykKTEHlhIXqrQ4kA3oo3rvywTN0Cc 2022-10-21 13:36:14 2022-10-21 13:36:14
## 2  fknPuv0Ub7Vj3yRQcKPy1YE9qERB6epJ 2022-10-21 13:36:57 2022-10-21 13:36:57
## 3  LWxWsT4BJ5RjQgTlM04nrKumcTnsLLJS 2022-10-21 13:40:11 2022-10-21 13:40:11
## 4  e0R1czs0WtY0N2WkCgMtyPjK7nvObN4j 2022-11-02 15:20:05 2022-11-02 15:20:05
## 5  0MI5CkV2sK5sQqN0Ywfq8m0YXzWDSg6b 2022-11-02 15:20:27 2022-11-02 15:20:27
## 6  CVmpcsTH0JzC8hWACgSu5YCjxS8xoiOv 2022-11-02 15:21:33 2022-11-02 15:21:33
## 7  hjy6Gs5Y5ypXjvMi36kC0rrWNhkn2bMM 2022-11-02 15:25:15 2022-11-02 15:25:15
## 8  ycOkcSSek4JedE89cJJFo1sDLnSI4iD1 2022-11-02 15:29:13 2022-11-02 15:29:13
## 9  RmvYY0jBSIASTwXD0IG9tibw64O8Ep0N 2022-11-02 15:30:10 2022-11-02 15:30:10
## 10 jsQSqwIr7uJluUaSwyYUtkP3scFkHjG2 2022-11-02 15:30:50 2022-11-02 15:30:50
## 11 7ez3WvOYrCCcOESpDUnfccgpTAKoHbas 2022-11-02 15:31:39 2022-11-02 15:31:39
## 12 bKtpL2W9Q9C2K8lzJOGCTYn9ssfAqxpO 2022-11-02 15:35:40 2022-11-02 15:35:40
## 13 emXyqarqxE5ct103QjVIqxvSnkFwemKJ 2022-11-02 15:36:19 2022-11-02 15:36:19
## 14 HyFijUlSP1Mp5dia4Ti1RfNwIueVb7sb 2022-11-02 15:37:03 2022-11-02 15:37:03
## 15 dTts3MX52t6sC1atACrJbYqO13tIPVLq 2022-11-02 15:38:20 2022-11-02 15:38:20
## 16 9p1JlBBeRVE1FrozVChzcv0jMOYUkeJ2 2022-11-02 15:39:24 2022-11-02 15:39:24
## 17 0kFPxYuqV76x0j2H2sldt0J6YamSYuuX 2022-11-02 15:25:38 2022-11-02 15:25:38
## 18 vsM6F1vB2JryWWvvuRTDS7K5uP5rwtnp 2023-02-09 19:42:36 2023-02-09 19:42:36
## 19 01ywOvoAHBuTzh1s0uKxquEUEbTEy3kO 2023-02-09 19:43:30 2023-02-09 19:43:30
## 20 wrtuKPOJgNswVJSJ3R6RIsOUUm7Umq7W 2023-02-16 21:37:29 2023-02-16 21:37:29
## 21 0VsADwwTXDXcrB2lqkHlip4Is5mBoghE 2023-04-20 18:52:56 2023-04-20 18:53:16
## 22 RE6nwOhKbjj8VUCSuBeH8pzj752k7itx 2023-05-02 16:18:49 2023-05-02 16:18:49
## 23 CWKliw4i1nJ62puTTwOa3vwxm2pTHcFQ 2023-05-15 15:42:38 2023-05-15 15:42:38
## 24 V6y0hiVGaCLUmDSLbvCjsz80tfnLb7Xj 2023-06-08 21:55:11 2023-06-08 21:55:11
## 25 g4et5FGxT6c5Qt1cR8kuQaGxMIE86L6Y 2023-08-09 19:05:42 2023-08-09 19:05:42
## 26 gfCAxK3EJQqaRfKM7tUjiIW5GaMxhIDu 2023-08-09 19:07:48 2023-08-09 19:07:48
##    core_product core_client
## 1          <NA>        GRAY
## 2          <NA>        STEL
## 3          <NA>        FPET
## 4          AGEM        NACM
## 5          SBPR        NACM
## 6          <NA>        BALM
## 7          CENT        AFFT
## 8          <NA>        CSUG
## 9          <NA>        CALC
## 10         <NA>        KUGO
## 11         <NA>        LUSB
```

```
## 12         <NA>        BARK
## 13         <NA>        ELMN
## 14         <NA>        SLIM
## 15         <NA>        SUNS
## 16         <NA>        WATR
## 17         BORA        AFFT
## 18         NAIL        ARCA
## 19         FUNG        ARCA
## 20         HUMF        HITS
## 21         HYIM        HITS
## 22         <NA>        IOVA
## 23         PACI        MINT
## 24         CRUN        CRUN
## 25                     OPOS
## 26         <NA>        OPOS
```

```
# Establish a connection to the database
con_2 <- dbConnect(RPostgres::Postgres(),
                   dbname = "rfi",
                   port = 5432,
                   user = "nycfmexp",
                   password = "Zxcv1234",
                   host = "rfi-production-cluster.cluster-cqx4czau66cg.us-east-1.rds.amazonaws.com")
```

```
dbListFields(con_2, "web_visit_results")
```

```
##  [1] "id"                "core_client"       "project_id"
##  [4] "user_id"           "session_id"        "unique_key"
##  [7] "postal_code"       "region"            "dma"
## [10] "dma_code"          "city"              "country"
## [13] "browser"           "device"            "device_type"
## [16] "search_engine"     "medium"            "source"
## [19] "platform"          "platform_version"  "bounce"
## [22] "session_date_time" "ip_address"        "pages"
## [25] "search_terms"      "language"          "latitude"
## [28] "longitude"         "organization"      "referrer"
## [31] "session_length"    "created_at"        "updated_at"
## [34] "core_product"      "last_access"
```

```
querynew <- dbGetQuery(con_2, "SELECT
                               SPLIT_PART(dma, ',', 1) AS city,
                               TRIM(SPLIT_PART(dma, ',', 2)) AS state,
                               CASE
                                 WHEN TRIM(SPLIT_PART(dma, ',', 2)) IN ('CT', 'MA', 'ME', 'NH', 'NY', 'I
                                 WHEN TRIM(SPLIT_PART(dma, ',', 2)) IN ('AZ', 'CA', 'NM', 'NV', 'UT') TH
                                 ELSE 'Unknown'
                               END AS region
                              FROM web_visit_results
                              WHERE medium = 'cpc'
                              AND TRIM(SPLIT_PART(dma, ',', 2)) IN ('CT', 'MA', 'ME', 'NH', 'NY', 'RI',


head(querynew)
```

```
##          city state    region
## 1    New York    NY Northeast
```

```
## 2   San Diego    CA Southwest
## 3 Albuquerque    NM Southwest
## 4      Boston    MA Northeast
## 5     Buffalo    NY Northeast
## 6    New York    NY Northeast
```

```
library(dplyr)

top_cities_northeast <- querynew %>%
  filter(region == 'Northeast') %>%
  group_by(city, state) %>%
  summarise(total_visits = n()) %>%
  arrange(desc(total_visits)) %>%
  top_n(10)
```

```
## `summarise()` has grouped output by 'city'. You can override using the
## `.groups` argument.
## Selecting by total_visits
```

```
top_cities_northeast
```

```
## # A tibble: 11 x 3
## # Groups:   city [11]
##    city                    state total_visits
##    <chr>                   <chr>        <int>
##  1 New York                NY              30
##  2 Boston                  MA              11
##  3 Albany-Schenectady-Troy NY               4
##  4 Buffalo                 NY               4
##  5 Hartford                CT               4
##  6 Syracuse                NY               3
##  7 Providence              RI               2
##  8 Portland-Auburn         ME               1
##  9 Rochester               NY               1
## 10 Springfield-Holyoke     MA               1
## 11 Utica-Rome              NY               1
```

```
library(dplyr)

top_cities_southwest <- querynew %>%
  filter(region == 'Southwest') %>%
  group_by(city, state) %>%
  summarise(total_visits = n()) %>%
  arrange(desc(total_visits)) %>%
  top_n(10)
```

```
## `summarise()` has grouped output by 'city'. You can override using the
## `.groups` argument.
## Selecting by total_visits
```

```
top_cities_southwest
```

```
## # A tibble: 11 x 3
## # Groups:   city [11]
##    city             state total_visits
##    <chr>            <chr>        <int>
##  1 Los Angeles      CA              14
```

```
##  2 San Francisco    CA               9
##  3 San Diego        CA               4
##  4 Monterey-Salinas CA               2
##  5 Phoenix          AZ               2
##  6 Sacramento       CA               2
##  7 Albuquerque      NM               1
##  8 Fresno           CA               1
##  9 Las Vegas        NV               1
## 10 Salt Lake City   UT               1
## 11 Santa Barbara    CA               1
```

```
querynew2 <- dbGetQuery(con_2, "SELECT
                                SPLIT_PART(dma, ',', 1) AS city,
                                TRIM(SPLIT_PART(dma, ',', 2)) AS state,
                                EXTRACT(MONTH FROM session_date_time) AS month,
                                CASE
                                  WHEN TRIM(SPLIT_PART(dma, ',', 2)) IN ('CT', 'MA', 'ME', 'NH', 'NY', '
                                  WHEN TRIM(SPLIT_PART(dma, ',', 2)) IN ('AZ', 'CA', 'NM', 'NV', 'UT') TH
                                  ELSE 'Unknown'
                                END AS region
                               FROM web_visit_results
                               WHERE medium = 'cpc'
                               AND TRIM(SPLIT_PART(dma, ',', 2)) IN ('CT', 'MA', 'ME', 'NH', 'NY', 'RI',

head(querynew2)
```

```
##            city state month    region
## 1      New York    NY     7 Northeast
## 2 Santa Barbara    CA     7 Southwest
## 3       Buffalo    NY     7 Northeast
## 4      New York    NY     7 Northeast
## 5        Boston    MA     7 Northeast
## 6        Boston    MA     7 Northeast
```

Northeast

```
library(dplyr)
library(knitr)
top_cities_northeast1 <- querynew2 %>%
  filter(region == 'Northeast') %>%
  group_by(city, state, month) %>%
  summarise(total_visits = n()) %>%
  filter(month %in% c(2, 3, 4, 5)) %>%
  arrange(month, desc(total_visits))
```

```
## `summarise()` has grouped output by 'city', 'state'. You can override using the
## `.groups` argument.
```

```
kable(top_cities_northeast1)
```

city	state	month	total_visits
New York	NY	2	28669
Boston	MA	2	9644
Hartford	CT	2	4078

city	state	month	total_visits
Buffalo	NY	2	2411
Providence	RI	2	2265
Albany-Schenectady-Troy	NY	2	1878
Rochester	NY	2	1310
Portland-Auburn	ME	2	1283
Springfield-Holyoke	MA	2	1056
Syracuse	NY	2	936
Binghamton	NY	2	449
Utica-Rome	NY	2	433
Elmira	NY	2	225
Watertown	NY	2	215
Bangor	ME	2	212
Presque Isle	ME	2	64
New York	NY	3	36532
Boston	MA	3	12507
Hartford	CT	3	5138
Buffalo	NY	3	3114
Providence	RI	3	2838
Albany-Schenectady-Troy	NY	3	2272
Portland-Auburn	ME	3	1590
Rochester	NY	3	1490
Springfield-Holyoke	MA	3	1234
Syracuse	NY	3	1089
Utica-Rome	NY	3	570
Binghamton	NY	3	466
Bangor	ME	3	286
Elmira	NY	3	270
Watertown	NY	3	245
Presque Isle	ME	3	70
New York	NY	4	41224
Boston	MA	4	14853
Hartford	CT	4	6314
Buffalo	NY	4	3768
Providence	RI	4	3519
Albany-Schenectady-Troy	NY	4	3163
Rochester	NY	4	2071
Portland-Auburn	ME	4	1853
Springfield-Holyoke	MA	4	1534
Syracuse	NY	4	1469
Utica-Rome	NY	4	700
Binghamton	NY	4	634
Elmira	NY	4	395
Watertown	NY	4	340
Bangor	ME	4	337
Presque Isle	ME	4	101
New York	NY	5	50914
Boston	MA	5	18733
Hartford	CT	5	8003
Buffalo	NY	5	4594
Providence	RI	5	4349
Albany-Schenectady-Troy	NY	5	3775
Rochester	NY	5	2409

city	state	month	total_visits
Portland-Auburn	ME	5	2375
Syracuse	NY	5	1897
Springfield-Holyoke	MA	5	1825
Utica-Rome	NY	5	968
Binghamton	NY	5	797
Bangor	ME	5	447
Elmira	NY	5	445
Watertown	NY	5	388
Presque Isle	ME	5	115

Percent Change For Northeast

```
library(dplyr)
library(knitr)

top_cities_northeast2 <- querynew2 %>%
  filter(region == 'Northeast') %>%
  group_by(city, state, month) %>%
  summarise(total_visits = n()) %>%
  filter(month %in% c(2, 3, 4, 5)) %>%
  arrange(city, month) %>%
  group_by(city) %>%
  mutate(percentage_change = (total_visits - lag(total_visits, default = first(total_visits))) / lag(to
  mutate(percentage_change = ifelse(percentage_change == 0, NA, percentage_change))
```

```
## `summarise()` has grouped output by 'city', 'state'. You can override using the
## `.groups` argument.
```

```
kable(top_cities_northeast2)
```

city	state	month	total_visits	percentage_change
Albany-Schenectady-Troy	NY	2	1878	NA
Albany-Schenectady-Troy	NY	3	2272	20.979766
Albany-Schenectady-Troy	NY	4	3163	39.216549
Albany-Schenectady-Troy	NY	5	3775	19.348720
Bangor	ME	2	212	NA
Bangor	ME	3	286	34.905660
Bangor	ME	4	337	17.832168
Bangor	ME	5	447	32.640950
Binghamton	NY	2	449	NA
Binghamton	NY	3	466	3.786192
Binghamton	NY	4	634	36.051502
Binghamton	NY	5	797	25.709779
Boston	MA	2	9644	NA
Boston	MA	3	12507	29.686852
Boston	MA	4	14853	18.757496
Boston	MA	5	18733	26.122669
Buffalo	NY	2	2411	NA
Buffalo	NY	3	3114	29.158026
Buffalo	NY	4	3768	21.001927
Buffalo	NY	5	4594	21.921444

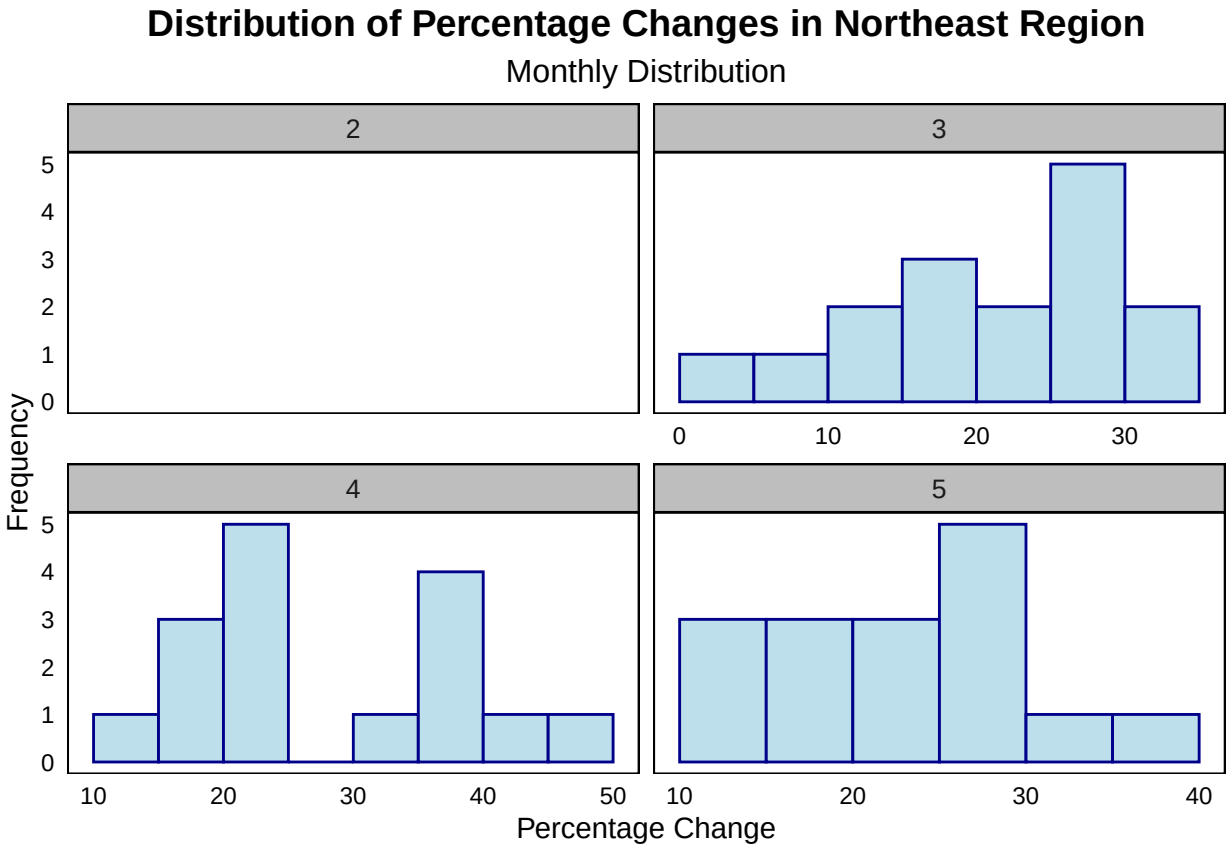
city	state	month	total_visits	percentage_change
Elmira	NY	2	225	NA
Elmira	NY	3	270	20.000000
Elmira	NY	4	395	46.296296
Elmira	NY	5	445	12.658228
Hartford	CT	2	4078	NA
Hartford	CT	3	5138	25.993134
Hartford	CT	4	6314	22.888283
Hartford	CT	5	8003	26.750079
New York	NY	2	28669	NA
New York	NY	3	36532	27.426837
New York	NY	4	41224	12.843534
New York	NY	5	50914	23.505725
Portland-Auburn	ME	2	1283	NA
Portland-Auburn	ME	3	1590	23.928293
Portland-Auburn	ME	4	1853	16.540881
Portland-Auburn	ME	5	2375	28.170534
Presque Isle	ME	2	64	NA
Presque Isle	ME	3	70	9.375000
Presque Isle	ME	4	101	44.285714
Presque Isle	ME	5	115	13.861386
Providence	RI	2	2265	NA
Providence	RI	3	2838	25.298013
Providence	RI	4	3519	23.995772
Providence	RI	5	4349	23.586246
Rochester	NY	2	1310	NA
Rochester	NY	3	1490	13.740458
Rochester	NY	4	2071	38.993289
Rochester	NY	5	2409	16.320618
Springfield-Holyoke	MA	2	1056	NA
Springfield-Holyoke	MA	3	1234	16.856061
Springfield-Holyoke	MA	4	1534	24.311183
Springfield-Holyoke	MA	5	1825	18.970013
Syracuse	NY	2	936	NA
Syracuse	NY	3	1089	16.346154
Syracuse	NY	4	1469	34.894399
Syracuse	NY	5	1897	29.135466
Utica-Rome	NY	2	433	NA
Utica-Rome	NY	3	570	31.639723
Utica-Rome	NY	4	700	22.807018
Utica-Rome	NY	5	968	38.285714
Watertown	NY	2	215	NA
Watertown	NY	3	245	13.953488
Watertown	NY	4	340	38.775510
Watertown	NY	5	388	14.117647

Histogram of percentage changes for each month in the Northeast region

```
ggplot(top_cities_northeast2, aes(x = percentage_change)) +
  geom_histogram(
    binwidth = 5,
    fill = "lightblue",
```

```
    color = "darkblue",
    alpha = 0.8,
    boundary = 0,
    show.legend = FALSE
  ) +
  facet_wrap(~ month, nrow = 2, scales = "free_x") +
  labs(
    x = "Percentage Change",
    y = "Frequency",
    title = "Distribution of Percentage Changes in Northeast Region",
    subtitle = "Monthly Distribution"
  ) +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.text = element_text(color = "black"),
    axis.title = element_text(color = "black"),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5),
    strip.text = element_text(size = 10),
    strip.background = element_rect(fill = "grey"),
    panel.background = element_rect(fill = "white")
  )
```

```
## Warning: Removed 16 rows containing non-finite values (`stat_bin()`).
```



Southwest

```
top_cities_southwest1 <- querynew2 %>%
  filter(region == 'Southwest') %>%
  group_by(city, state, month) %>%
  summarise(total_visits = n()) %>%
  filter(month %in% c(2, 3, 4, 5)) %>%
  arrange(month, desc(total_visits))
```

```
## `summarise()` has grouped output by 'city', 'state'. You can override using the
## `.groups` argument.
```

```
top_cities_southwest1
```

```
## # A tibble: 72 x 4
## # Groups:   city, state [18]
##    city           state month total_visits
##    <chr>          <chr> <dbl>        <int>
##  1 Los Angeles    CA        2        17463
##  2 San Francisco  CA        2         7532
##  3 Phoenix        AZ        2         5464
##  4 Sacramento     CA        2         4296
##  5 San Diego      CA        2         2691
##  6 Salt Lake City UT        2         2500
##  7 Las Vegas      NV        2         2343
##  8 Fresno         CA        2         1157
##  9 Albuquerque    NM        2          985
## 10 Tucson         AZ        2          968
## # i 62 more rows
```

```
kable(top_cities_southwest1)
```

city	state	month	total_visits
Los Angeles	CA	2	17463
San Francisco	CA	2	7532
Phoenix	AZ	2	5464
Sacramento	CA	2	4296
San Diego	CA	2	2691
Salt Lake City	UT	2	2500
Las Vegas	NV	2	2343
Fresno	CA	2	1157
Albuquerque	NM	2	985
Tucson	AZ	2	968
Reno	NV	2	789
Palm Springs	CA	2	650
Santa Barbara	CA	2	585
Monterey-Salinas	CA	2	529
Bakersfield	CA	2	476
Chico-Redding	CA	2	457
Yuma	AZ	2	269
Eureka	CA	2	122
Los Angeles	CA	3	19113
San Francisco	CA	3	8753
Phoenix	AZ	3	5793
Sacramento	CA	3	4520

city	state	month	total_visits
San Diego	CA	3	2856
Salt Lake City	UT	3	2773
Las Vegas	NV	3	2407
Fresno	CA	3	1383
Albuquerque	NM	3	994
Tucson	AZ	3	990
Reno	NV	3	939
Santa Barbara	CA	3	614
Palm Springs	CA	3	601
Monterey-Salinas	CA	3	520
Bakersfield	CA	3	515
Chico-Redding	CA	3	441
Yuma	AZ	3	266
Eureka	CA	3	111
Los Angeles	CA	4	20030
San Francisco	CA	4	8955
Phoenix	AZ	4	6589
Sacramento	CA	4	5062
Salt Lake City	UT	4	3372
San Diego	CA	4	3047
Las Vegas	NV	4	2605
Fresno	CA	4	1359
Tucson	AZ	4	1264
Albuquerque	NM	4	1170
Reno	NV	4	1005
Palm Springs	CA	4	734
Santa Barbara	CA	4	622
Monterey-Salinas	CA	4	608
Bakersfield	CA	4	575
Chico-Redding	CA	4	507
Yuma	AZ	4	302
Eureka	CA	4	135
Los Angeles	CA	5	21881
San Francisco	CA	5	10489
Phoenix	AZ	5	7109
Sacramento	CA	5	6094
Salt Lake City	UT	5	4032
San Diego	CA	5	3219
Las Vegas	NV	5	2910
Fresno	CA	5	1593
Albuquerque	NM	5	1394
Reno	NV	5	1274
Tucson	AZ	5	1235
Palm Springs	CA	5	750
Santa Barbara	CA	5	736
Bakersfield	CA	5	675
Monterey-Salinas	CA	5	672
Chico-Redding	CA	5	646
Yuma	AZ	5	313
Eureka	CA	5	158

Percent Change for Southwest

```
library(dplyr)

top_cities_southwest2 <- querynew2 %>%
  filter(region == 'Southwest') %>%
  group_by(city, state, month) %>%
  summarise(total_visits = n()) %>%
  filter(month %in% c(2, 3, 4, 5)) %>%
  arrange(city, month) %>%
  group_by(city) %>%
  mutate(percentage_change = (total_visits - lag(total_visits,
  default = first(total_visits))) / lag(total_visits, default = first(total_visits)) * 100) %>%
   mutate(percentage_change = ifelse(percentage_change == 0, NA, percentage_change))
```

```
## `summarise()` has grouped output by 'city', 'state'. You can override using the
## `.groups` argument.
```

```
kable(top_cities_southwest2)
```

city	state	month	total_visits	percentage_change
Albuquerque	NM	2	985	NA
Albuquerque	NM	3	994	0.9137056
Albuquerque	NM	4	1170	17.7062374
Albuquerque	NM	5	1394	19.1452991
Bakersfield	CA	2	476	NA
Bakersfield	CA	3	515	8.1932773
Bakersfield	CA	4	575	11.6504854
Bakersfield	CA	5	675	17.3913043
Chico-Redding	CA	2	457	NA
Chico-Redding	CA	3	441	-3.5010941
Chico-Redding	CA	4	507	14.9659864
Chico-Redding	CA	5	646	27.4161736
Eureka	CA	2	122	NA
Eureka	CA	3	111	-9.0163934
Eureka	CA	4	135	21.6216216
Eureka	CA	5	158	17.0370370
Fresno	CA	2	1157	NA
Fresno	CA	3	1383	19.5332757
Fresno	CA	4	1359	-1.7353579
Fresno	CA	5	1593	17.2185430
Las Vegas	NV	2	2343	NA
Las Vegas	NV	3	2407	2.7315408
Las Vegas	NV	4	2605	8.2260075
Las Vegas	NV	5	2910	11.7082534
Los Angeles	CA	2	17463	NA
Los Angeles	CA	3	19113	9.4485484
Los Angeles	CA	4	20030	4.7977816
Los Angeles	CA	5	21881	9.2411383
Monterey-Salinas	CA	2	529	NA
Monterey-Salinas	CA	3	520	-1.7013233
Monterey-Salinas	CA	4	608	16.9230769
Monterey-Salinas	CA	5	672	10.5263158
Palm Springs	CA	2	650	NA

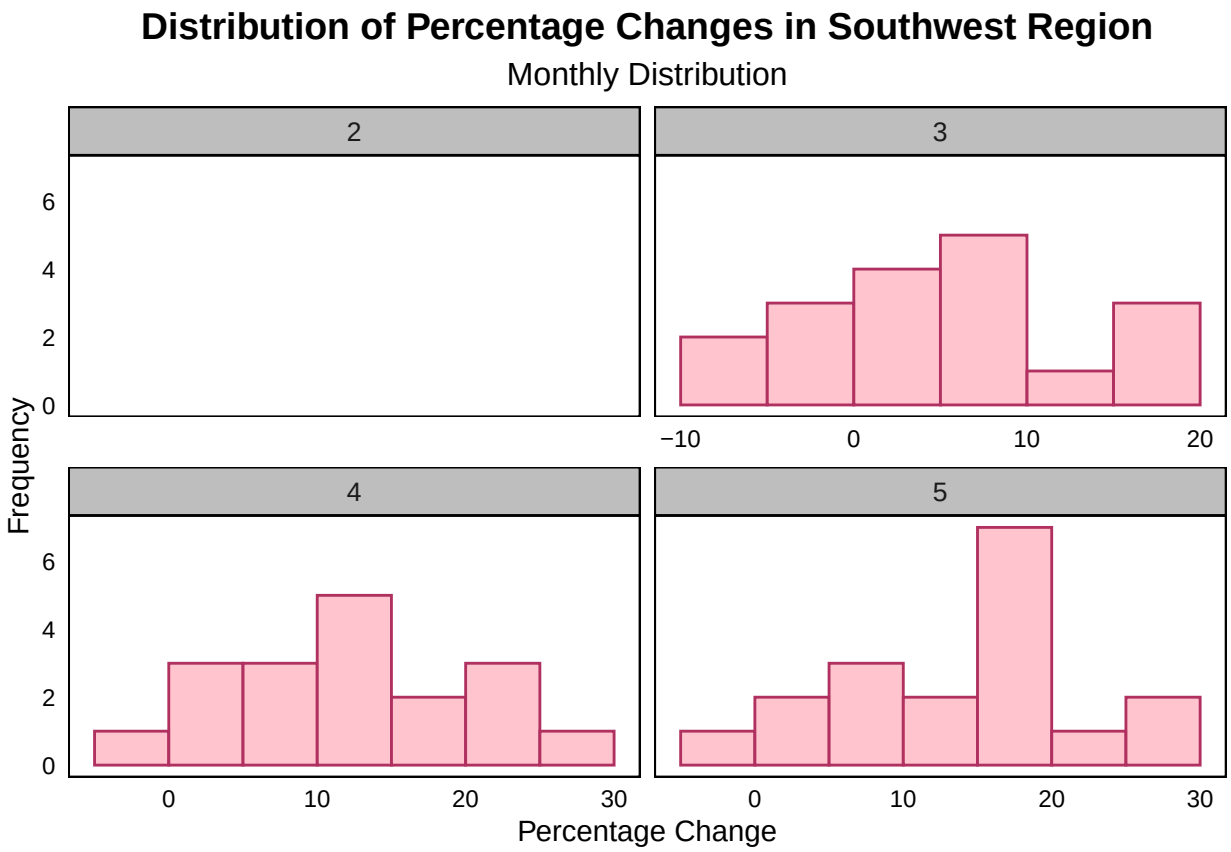
city	state	month	total_visits	percentage_change
Palm Springs	CA	3	601	-7.5384615
Palm Springs	CA	4	734	22.1297837
Palm Springs	CA	5	750	2.1798365
Phoenix	AZ	2	5464	NA
Phoenix	AZ	3	5793	6.0212299
Phoenix	AZ	4	6589	13.7407216
Phoenix	AZ	5	7109	7.8919411
Reno	NV	2	789	NA
Reno	NV	3	939	19.0114068
Reno	NV	4	1005	7.0287540
Reno	NV	5	1274	26.7661692
Sacramento	CA	2	4296	NA
Sacramento	CA	3	4520	5.2141527
Sacramento	CA	4	5062	11.9911504
Sacramento	CA	5	6094	20.3871987
Salt Lake City	UT	2	2500	NA
Salt Lake City	UT	3	2773	10.9200000
Salt Lake City	UT	4	3372	21.6011540
Salt Lake City	UT	5	4032	19.5729537
San Diego	CA	2	2691	NA
San Diego	CA	3	2856	6.1315496
San Diego	CA	4	3047	6.6876751
San Diego	CA	5	3219	5.6448966
San Francisco	CA	2	7532	NA
San Francisco	CA	3	8753	16.2108338
San Francisco	CA	4	8955	2.3077802
San Francisco	CA	5	10489	17.1300949
Santa Barbara	CA	2	585	NA
Santa Barbara	CA	3	614	4.9572650
Santa Barbara	CA	4	622	1.3029316
Santa Barbara	CA	5	736	18.3279743
Tucson	AZ	2	968	NA
Tucson	AZ	3	990	2.2727273
Tucson	AZ	4	1264	27.6767677
Tucson	AZ	5	1235	-2.2943038
Yuma	AZ	2	269	NA
Yuma	AZ	3	266	-1.1152416
Yuma	AZ	4	302	13.5338346
Yuma	AZ	5	313	3.6423841

Histogram of percentage changes for each month in the Southwest region

```
ggplot(top_cities_southwest2, aes(x = percentage_change)) +
  geom_histogram(
    binwidth = 5,
    fill = "lightpink",
    color = "maroon",
    alpha = 0.8,
    boundary = 0,
    show.legend = FALSE
  ) +
```

```
  facet_wrap(~ month, nrow = 2, scales = "free_x") +
  labs(
    x = "Percentage Change",
    y = "Frequency",
    title = "Distribution of Percentage Changes in Southwest Region",
    subtitle = "Monthly Distribution"
  ) +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.text = element_text(color = "black"),
    axis.title = element_text(color = "black"),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5),
    strip.text = element_text(size = 10),
    strip.background = element_rect(fill = "grey"),
    panel.background = element_rect(fill = "white")
  )
```

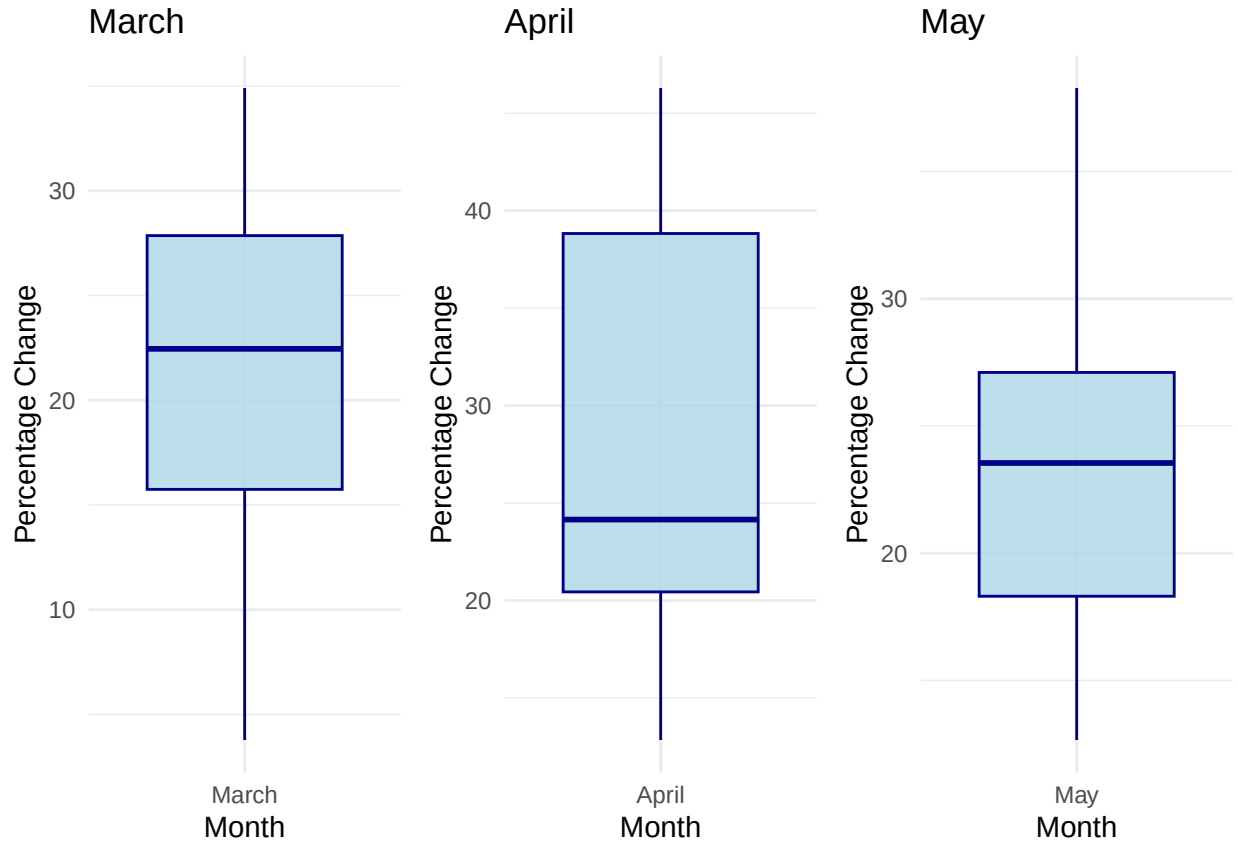
```
## Warning: Removed 18 rows containing non-finite values (`stat_bin()`).
```



Boxplot for Northeast

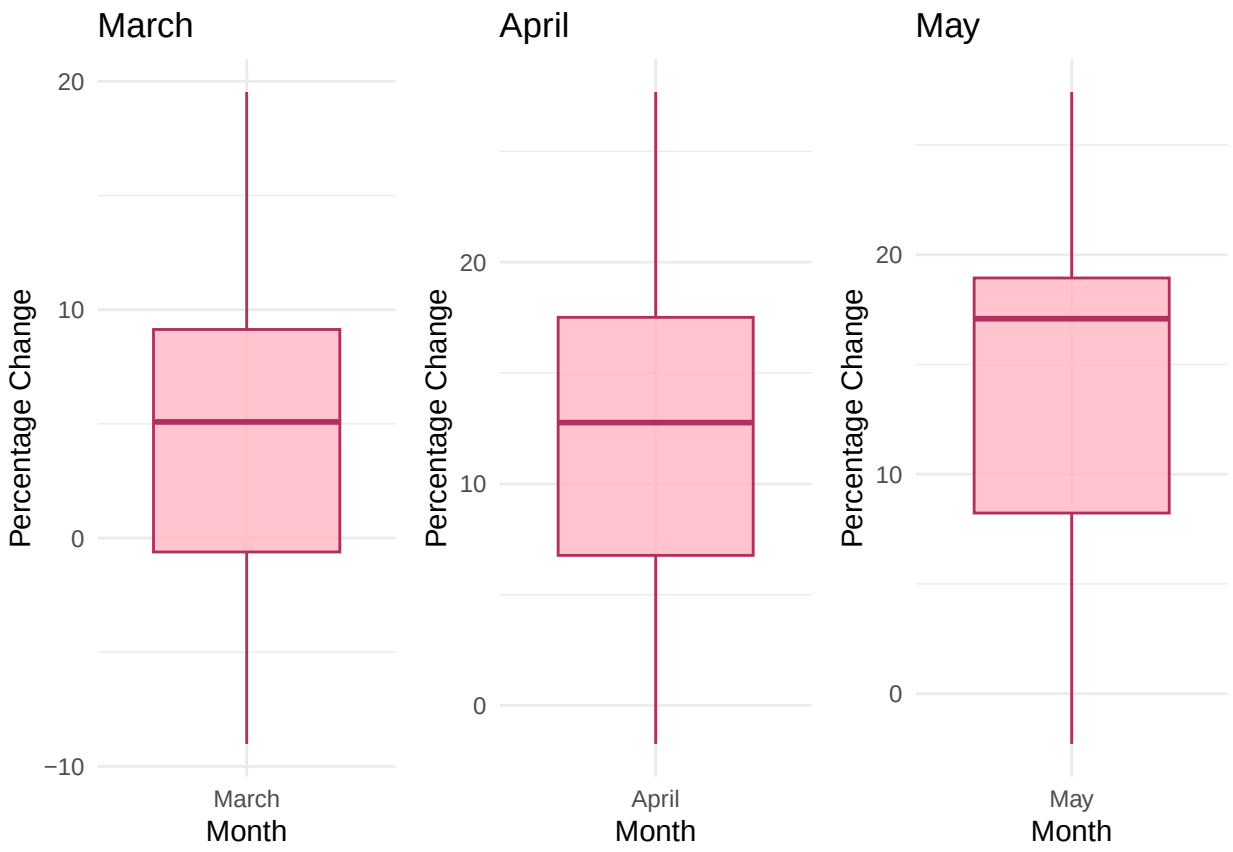
```
library(ggplot2)
```

```r
# Boxplot for Northeast region - March
p4 <- ggplot(top_cities_northeast2[top_cities_northeast2$month == 3, ], aes(x = "March", y = percentage_
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.8) +
  labs(title = "March",
       x = "Month",
       y = "Percentage Change") +
  theme_minimal()

# Boxplot for Northeast region - April
p5 <- ggplot(top_cities_northeast2[top_cities_northeast2$month == 4, ], aes(x = "April", y = percentage_
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.8) +
  labs(title = "April",
       x = "Month",
       y = "Percentage Change") +
  theme_minimal()

# Boxplot for Northeast region - May
p6 <- ggplot(top_cities_northeast2[top_cities_northeast2$month == 5, ], aes(x = "May", y = percentage_ch
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.8) +
  labs(title = "May",
       x = "Month",
       y = "Percentage Change") +
  theme_minimal()

# Combine the plots into a 1x3 grid
grid.arrange(p4, p5, p6, nrow = 1)
```



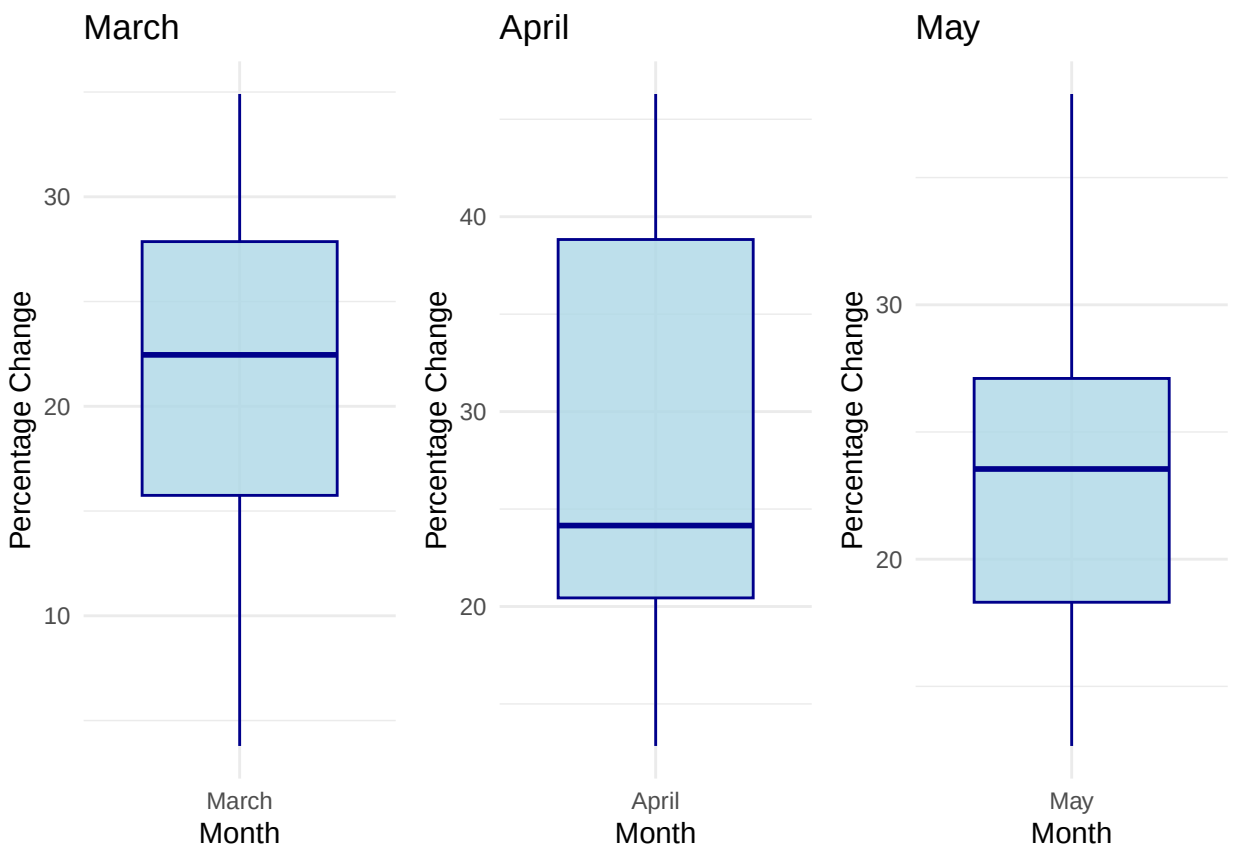
Boxplot for Southwest

```
library(ggplot2)
library(gridExtra)

# Boxplot for Southwest region - March
p1 <- ggplot(top_cities_southwest2[top_cities_southwest2$month == 3, ], aes(x = "March", y = percentage_
  geom_boxplot(fill = "lightpink", color = "maroon", alpha = 0.8) +
  labs(title = "March",
       x = "Month",
       y = "Percentage Change") +
  theme_minimal()

# Boxplot for Southwest region - April
p2 <- ggplot(top_cities_southwest2[top_cities_southwest2$month == 4, ], aes(x = "April", y = percentage_
  geom_boxplot(fill = "lightpink", color = "maroon", alpha = 0.8) +
  labs(title = "April",
       x = "Month",
       y = "Percentage Change") +
  theme_minimal()

# Boxplot for Southwest region - May
p3 <- ggplot(top_cities_southwest2[top_cities_southwest2$month == 5, ], aes(x = "May", y = percentage_ch
  geom_boxplot(fill = "lightpink", color = "maroon", alpha = 0.8) +
  labs(title = "May ",
       x = "Month",
       y = "Percentage Change") +
  theme_minimal()

# Combine the plots into a 1x3 grid
grid.arrange(p1, p2, p3, nrow = 1)
```

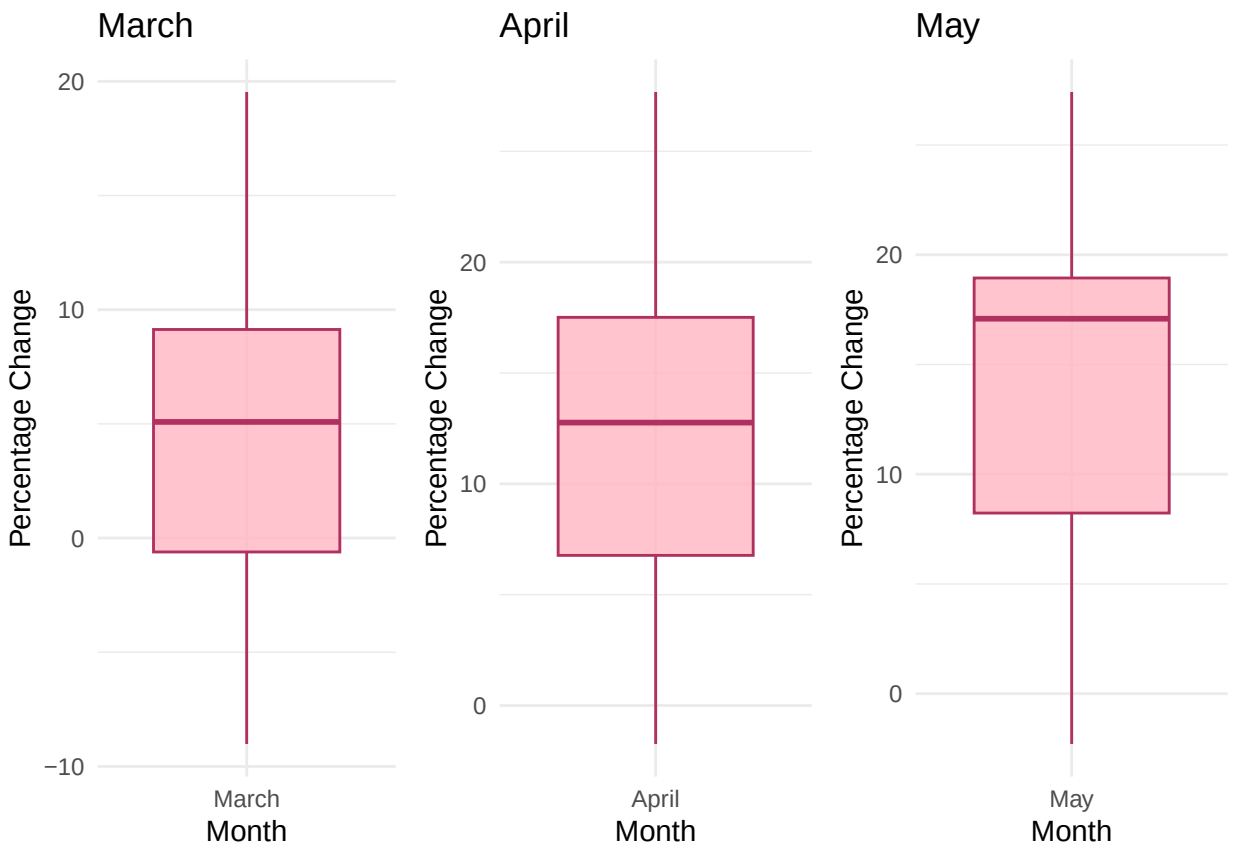


Comparision of Boxplot

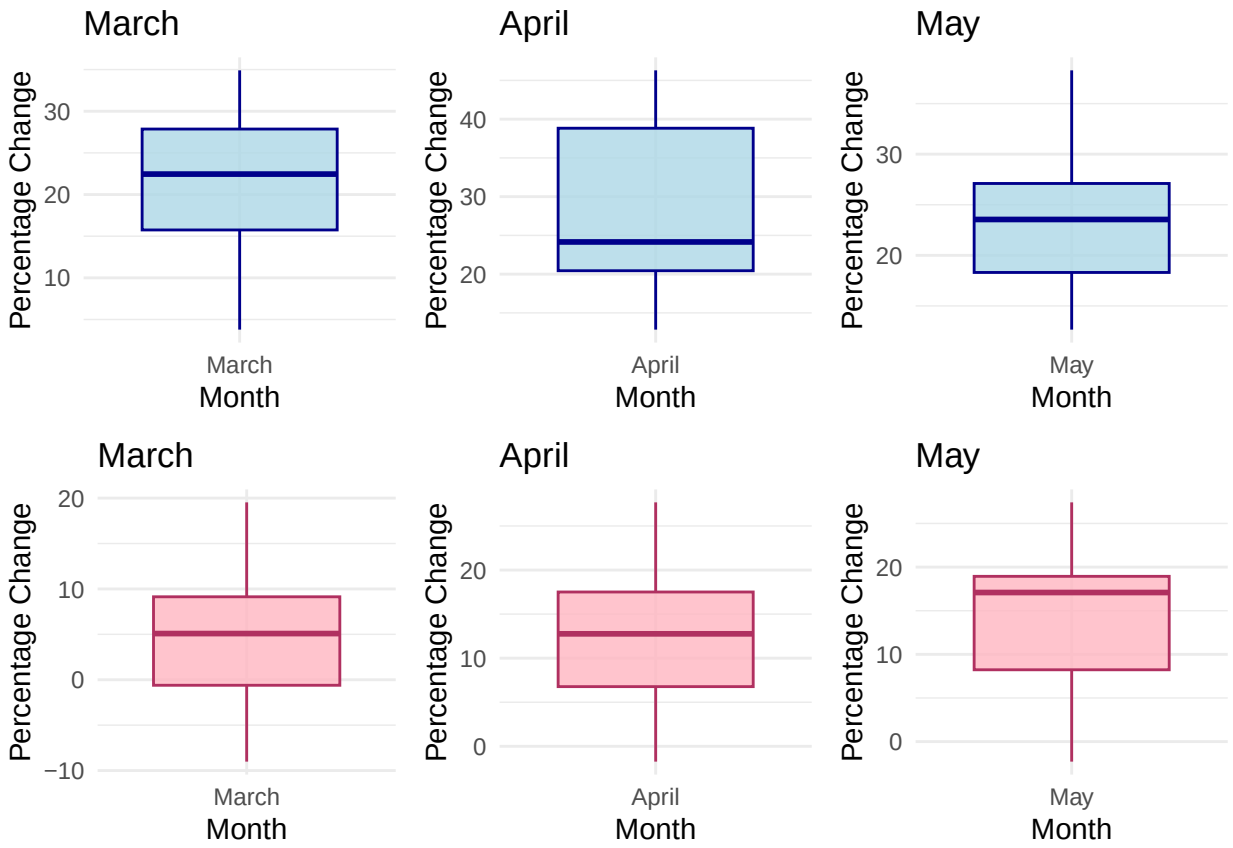
```
library(gridExtra)

# Combine the plots for Northeast region
northeast_plots <- grid.arrange(p4, p5, p6, nrow = 1)
```




```
# Combine the plots for Southwest region
southwest_plots <- grid.arrange(p1, p2, p3, nrow = 1)
```

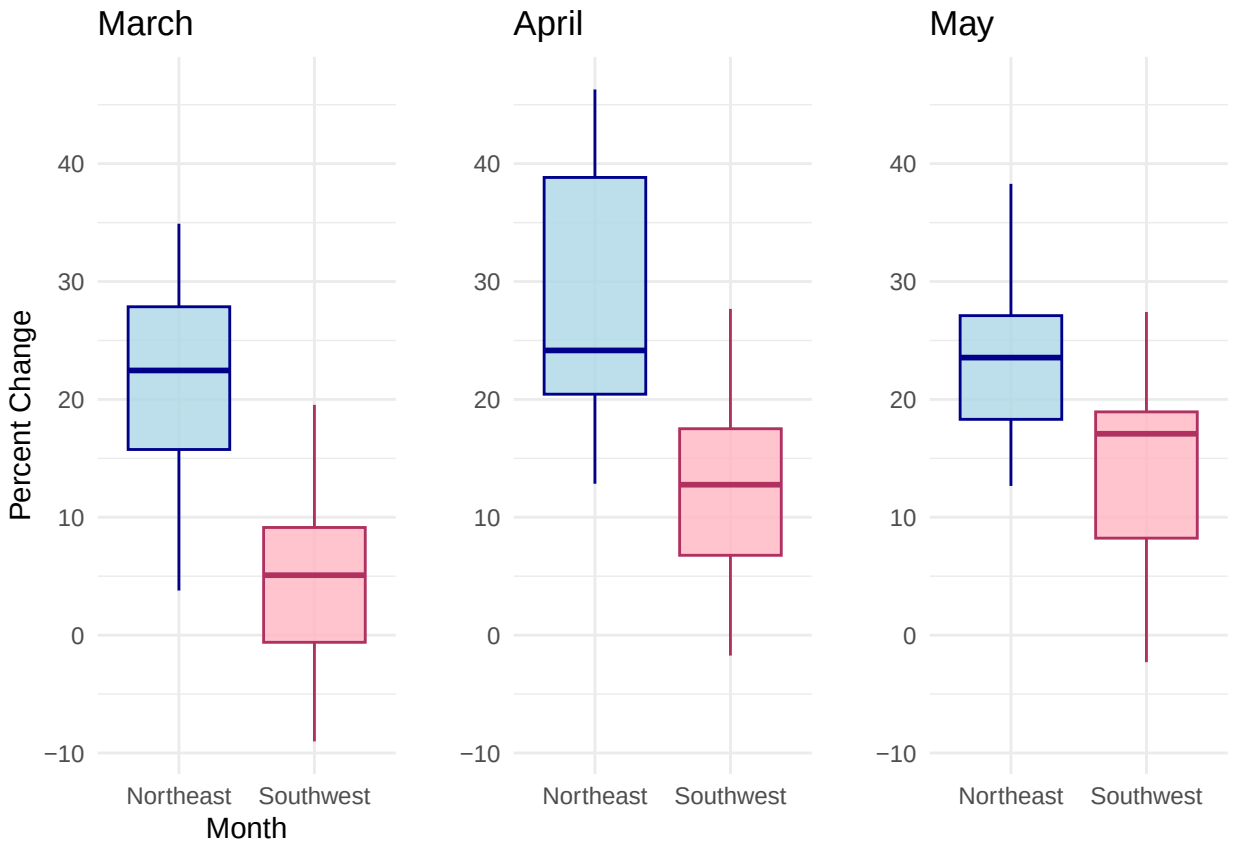



```
# Combine the plots for both regions
grid.arrange(northeast_plots, southwest_plots, nrow = 2)
```




```r
library(ggplot2)
library(gridExtra)

# Create a combined data frame for March
march_data <- rbind(
  transform(top_cities_northeast2[top_cities_northeast2$month == 3, ], region = "Northeast"),
  transform(top_cities_southwest2[top_cities_southwest2$month == 3, ], region = "Southwest")
)

# Create a combined data frame for April
april_data <- rbind(
  transform(top_cities_northeast2[top_cities_northeast2$month == 4, ], region = "Northeast"),
  transform(top_cities_southwest2[top_cities_southwest2$month == 4, ], region = "Southwest")
)

# Create a combined data frame for May
may_data <- rbind(
  transform(top_cities_northeast2[top_cities_northeast2$month == 5, ], region = "Northeast"),
  transform(top_cities_southwest2[top_cities_southwest2$month == 5, ], region = "Southwest")
)

# Calculate the minimum and maximum values for the y-axis scale
min_value <- min(min(march_data$percentage_change), min(april_data$percentage_change), min(may_data$per
max_value <- max(max(march_data$percentage_change), max(april_data$percentage_change), max(may_data$per

# Combine the box plots for each month
combined_plot <- grid.arrange(
```

```
  ggplot(march_data, aes(x = region, y = percentage_change)) +
    geom_boxplot(fill = c("lightblue", "lightpink"), color = c("darkblue", "maroon"), alpha = 0.8) +
    labs(title = "March", x = "", y = "Percent Change") +
    xlab("Month") +
    coord_cartesian(ylim = c(min_value, max_value)) +
    theme_minimal(),

  ggplot(april_data, aes(x = region, y = percentage_change)) +
    geom_boxplot(fill = c("lightblue", "lightpink"), color = c("darkblue", "maroon"), alpha = 0.8) +
    labs(title = "April", x = "", y = "") +
    coord_cartesian(ylim = c(min_value, max_value)) +
    theme_minimal(),

  ggplot(may_data, aes(x = region, y = percentage_change)) +
    geom_boxplot(fill = c("lightblue", "lightpink"), color = c("darkblue", "maroon"), alpha = 0.8) +
    labs(title = "May", x = "", y = "") +
    coord_cartesian(ylim = c(min_value, max_value)) +
    theme_minimal(),

  nrow = 1
)
```




```
# Display the combined plot
print(combined_plot)
```

```
## TableGrob (1 x 3) "arrange": 3 grobs
##   z     cells    name           grob
```

```
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## 3 3 (1-1,3-3) arrange gtable[layout]
```

Breakdown of what the graph means: The box: It represents the interquartile range (IQR) of the data, spanning from the first quartile (Q1) to the third quartile (Q3). The height of the box indicates the spread of the middle 50% of the data. The horizontal line within the box: It represents the median (Q2) of the data, which indicates the central tendency. The whiskers: They extend from the box to show the range of the data, excluding outliers. The whiskers typically extend 1.5 times the IQR from the box. Any data points beyond the whiskers are considered outliers and plotted as individual points. The outliers: Individual points outside the whiskers are plotted as outliers. They represent values that are significantly different from the rest of the data.

This Graph represents the percent change of visits for the Northeast and Southwest regions. Looking at the two regions side by side, it is evident that the Northeast Region has a higher percent change of visits compared to the southwest region for month by month. The data has been analyzed for the months of March, April, and May.In March, the box plot shows that the Northeast region had a higher percentage change of visits compared to the Southwest region. The box plot for April also demonstrates a similar trend, with the Northeast region showing a larger increase in visits. The same pattern continues in May, where the box plot again indicates a higher percentage change in visits for the Northeast region. These findings suggest that the Northeast region experiences a consistently higher increase in visits per month compared to the Southwest region. The box plots visually illustrate the distribution of the percentage changes, with the box representing the interquartile range and the whiskers indicating the range of the data. The difference in the box plot positions and whisker lengths between the two regions highlights the contrasting levels of visitation growth. In summary, the graph effectively conveys the higher percentage change of visits in the Northeast region compared to the Southwest region over the analyzed months.

A few Reasons to why this is happening:

1. Seasonal Factors: The higher percentage change of visits in the Northeast region compared to the Southwest region during the analyzed months (March, April, and May) could be influenced by the seasonal factor, where the arrival of spring leads to increased outdoor activities and a greater demand for awnings in the Northeast region after enduring a long winter, while the milder winters in the Southwest region result in a relatively lower seasonal shift in demand.
2. Competition and Market Dynamics: The varying levels of competition and market dynamics between the Northeast and Southwest regions, such as a potentially more saturated awning market in the Southwest with the presence of local competitors, pricing strategies, and consumer preferences, can contribute to the observed differences in the percent change of visits for Sunsetter.

Test One

Null hypothesis (H0): Percentage changes in search volume for the Northeast and Southwest regions are not meaningfully different.

Alternative hypothesis (H1): Percentage changes in search volume for the Northeast region are significantly different from the Southwest region.

```
# Subset the data for the Northeast and Southwest regions
percent_change_northeast <- top_cities_northeast2$percentage_change
percent_change_southwest <- top_cities_southwest2$percentage_change

# Perform an independent samples t-test
result <- t.test(percent_change_northeast, percent_change_southwest)

# Print the result
result
```

```
##
##  Welch Two Sample t-test
##
## data:  percent_change_northeast and percent_change_southwest
## t = 7.7628, df = 97.461, p-value = 8.335e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.48425 17.68622
## sample estimates:
## mean of x mean of y
##  24.45147  10.36623
```

The t-statistic computed was 7.7628, and the degrees of freedom were calculated as 97.461. The corresponding p-value obtained from the test was 8.335e-12, indicating an extremely small value. This suggests strong evidence against the null hypothesis. Here, 7.7628 is a high value, which suggests a large difference between the means of the northeast and southwest groups.

The p-value is extremely small (0.000000000008335), far below the common alpha level of 0.05, so we reject the null hypothesis. This means it's very unlikely the observed data would occur if there was no true difference in means.

A 95% confidence interval is a range of values that you can be 95% certain contains the true mean difference between the two population. This means we are 95% confident that the average percent change in the northeast is between 10.48425 and 17.68622 units larger than in the southwest.

Sample estimates: mean of x mean of y 24.45147 10.36623: These are the sample means for the two groups. The average percent change for the northeast (x) is 24.45147 and for the southwest (y) is 10.36623.

In conclusion, the results of the Welch Two Sample t-test provide compelling evidence to suggest that there is a significant and meaningful difference in the percentage changes in search volume between the Northeast and Southwest regions. The Northeast region shows a notably higher mean percentage change compared to the Southwest region, with the confidence interval further supporting this conclusion.

Test Two

Null hypothesis (H0): Percentage changes in search volume for the Northeast and Southwest regions are not meaningfully different.

Alternative hypothesis (H1): Percentage changes in search volume for the Northeast region are significantly different from the Southwest region.

```
# Load necessary packages
library(dplyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

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

```
library(knitr)

# Determine the number of rows
```

```r
num_rows <- nrow(top_cities_northeast2)

# Exclude top 3 and bottom 3 Northeast markets by search volume
top_cities_northeast2_filtered <- top_cities_northeast2 %>%
  arrange(desc(total_visits)) %>%
  tail(num_rows - 3) %>%
  head(num_rows - 6)

# Exclude top 3 and bottom 3 Southwest markets by search volume
top_cities_southwest2_filtered <- top_cities_southwest2 %>%
  arrange(desc(total_visits)) %>%
  tail(num_rows - 3) %>%
  head(num_rows - 6)

# Subset the data for the filtered Northeast and Southwest markets
percent_change_northeast_filtered <- top_cities_northeast2_filtered$percentage_change
percent_change_southwest_filtered <- top_cities_southwest2_filtered$percentage_change

# Perform an independent samples t-test
result2 <- t.test(percent_change_northeast_filtered, percent_change_southwest_filtered)

# Print the result
result2
```

```
##
##  Welch Two Sample t-test
##
## data:  percent_change_northeast_filtered and percent_change_southwest_filtered
## t = 7.335, df = 83.976, p-value = 1.278e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.34372 18.03852
## sample estimates:
## mean of x mean of y
##  24.56357  10.37245
```

The t-statistic calculated was 7.335, and the degrees of freedom were determined as 83.976. The corresponding p-value obtained from the test was 1.278e-10, which is extremely small. This provides substantial evidence against the null hypothesis. Here, 7.335 suggests that the means of the two groups are quite different.

In this case, the null hypothesis is that there is no difference between the two means. A small p-value (usually less than 0.05) tells us that the likelihood of observing the data given that the null hypothesis is true is very low. Therefore, we reject the null hypothesis and conclude that there is a significant difference between the two means.

95 percent confidence interval: 10.34372 18.03852: This is the 95% confidence interval for the difference between the two group means. We can be 95% confident that the true difference between the population means falls within this interval.

The sample estimates indicate that the mean percentage change in search volume for the Northeast region is 24.56357, while for the Southwest region, it is 10.37245.

In summary, this t-test suggests that there is a significant difference between the mean percent changes in the northeast and southwest. The northeast's mean percent change is statistically significantly higher than the southwest's, and we can be 95% confident that the true difference in means falls between 10.34 and 18.04.

Putting it all Together

In Test One:

The t-statistic is 7.7628, with a corresponding p-value of 8.335e-12. The 95% confidence interval for the difference in means is 10.48425 to 17.68622. The mean percentage change in search volume is 24.45147 for the Northeast region and 10.36623 for the Southwest region.

In Test Two:

The t-statistic is 7.335, with a p-value of 1.278e-10. The 95% confidence interval for the difference in means is 10.34372 to 18.03852. The mean percentage change in search volume is 24.56357 for the Northeast region and 10.37245 for the Southwest region. Comparing the two tests, we observe that excluding the top 3 and bottom 3 markets in Test Two results in slightly lower t-statistics and wider confidence intervals compared to Test One. However, the p-values remain extremely small in both tests, indicating strong evidence against the null hypothesis.

The mean percentage change in search volume for the Northeast region remains higher than that of the Southwest region in both tests.

Overall, the comparison suggests that even after excluding the top 3 and bottom 3 markets by search volume, the findings remain consistent. There is a significant and meaningful difference in the percentage changes in search volume between the Northeast and Southwest regions, with the Northeast region exhibiting higher mean percentage changes.

These results support the conclusion that the Northeast and Southwest regions differ significantly in terms of search volume percentage changes, regardless of the inclusion or exclusion of the top and bottom markets.

Business Interpretation

Given the statistical evidence showing a significant difference in search volume changes between the Northeast and Southwest regions, there are several ways this could be leveraged to make marketing efforts more efficient:

1. Seasonal Marketing: Since the increase in search volumes during the spring is higher in the Northeast region compared to the Southwest, we might consider adjusting our marketing campaigns to this seasonal trend. For instance, we can allocate more budget for spring campaigns in the Northeast to tap into the increased search volumes.
2. Optimizing Ad Spend: As search volumes in the Northeast increase significantly more than in the Southwest during spring, focusing more of our marketing efforts and spending in the Northeast during this time might yield higher returns.
3. Market Opportunities: The higher search volume in the Northeast indicates a potentially larger or more engaged market for your business or product. This could represent an opportunity to focus marketing and sales efforts in this region to capitalize on this interest.
4. Personalized Marketing: Using this insight, we can design and deliver more personalized marketing messages to different regions. For example, spring-themed marketing campaigns can be more heavily promoted in the Northeast region, while the Southwest region might respond better to a different seasonal focus.
5. Strategy Adjustment: The Southwest region might require different marketing strategies, products, or services to increase its search volume and match the performance of the Northeast region. Market research could be beneficial here to understand why the Southwest region is underperforming and how this could be addressed.