Tidy Time Series & Forecasting in R

5. Time series features

bit.ly/fable2020

Outline

1. STL Features
2. Lab Session 9
3. Dimension reduction for features
4. Lab Session 10

Outline

Strength of seasonality and trend

STL decomposition

$$y_t = T_t + S_t + R_t$$

Seasonal strength

$$\max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t + R_t)}\right)$$

Trend strength

$$\max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t + R_t)}\right)$$

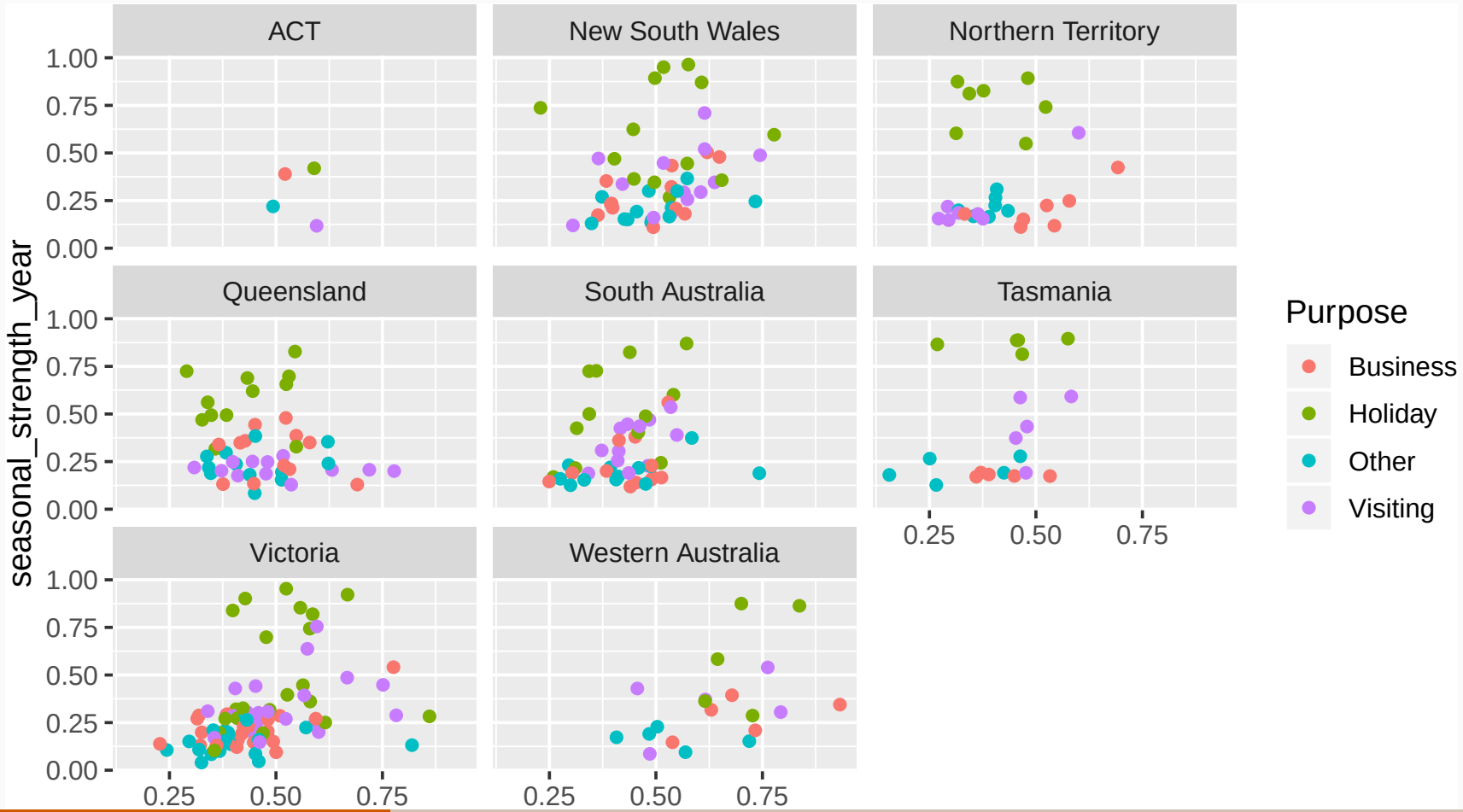
Feature extraction and statistics

```
tourism %>% features(Trips, feat_stl)
```

```
## # A tibble: 304 x 12
##    Region State Purpose trend_strength seasonal_streng~
##    <chr>  <chr> <chr>            <dbl>            <dbl>
##  1 Adela~ Sout~ Busine~          0.451            0.380
##  2 Adela~ Sout~ Holiday          0.541            0.601
##  3 Adela~ Sout~ Other            0.743            0.189
##  4 Adela~ Sout~ Visiti~          0.433            0.446
##  5 Adela~ Sout~ Busine~          0.453            0.140
##  6 Adela~ Sout~ Holiday          0.512            0.244
##  7 Adela~ Sout~ Other            0.584            0.374
##  8 Adela~ Sout~ Visiti~          0.481            0.228
##  9 Alice~ Nort~ Busine~          0.526            0.224
## 10 Alice~ Nort~ Holiday          0.377            0.827
## # ... with 294 more rows, and 7 more variables:
## #   seasonal_peak_year <dbl>, seasonal_trough_year <dbl>,
## #   spikiness <dbl>, linearity <dbl>, curvature <dbl>,
## #   stl_e_acf1 <dbl>, stl_e_acf10 <dbl>
```

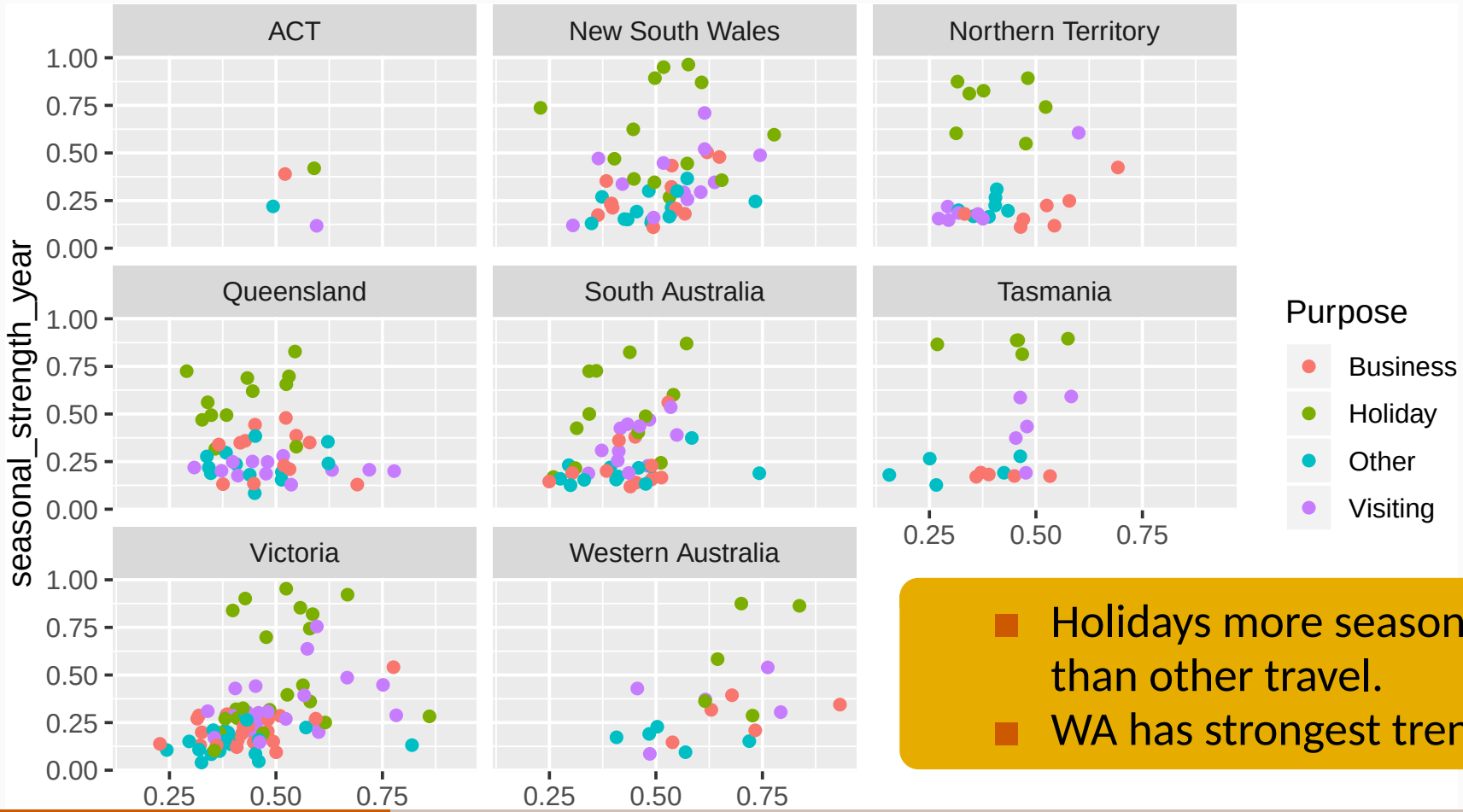
Feature extraction and statistics

```
tourism %>%
  features(Trips, feat_stl) %>%
  ggplot(aes(x = trend_strength, y = seasonal_strength_year, col = Purpose)) +
  geom_point() + facet_wrap(vars(State))
```



Feature extraction and statistics

```
tourism %>%
  features(Trips, feat_stl) %>%
  ggplot(aes(x = trend_strength, y = seasonal_strength_year, col = Purpose)) +
  geom_point() + facet_wrap(vars(State))
```



- Holidays more seasonal than other travel.
- WA has strongest trends.

Feature extraction and statistics

Find the most seasonal time series:

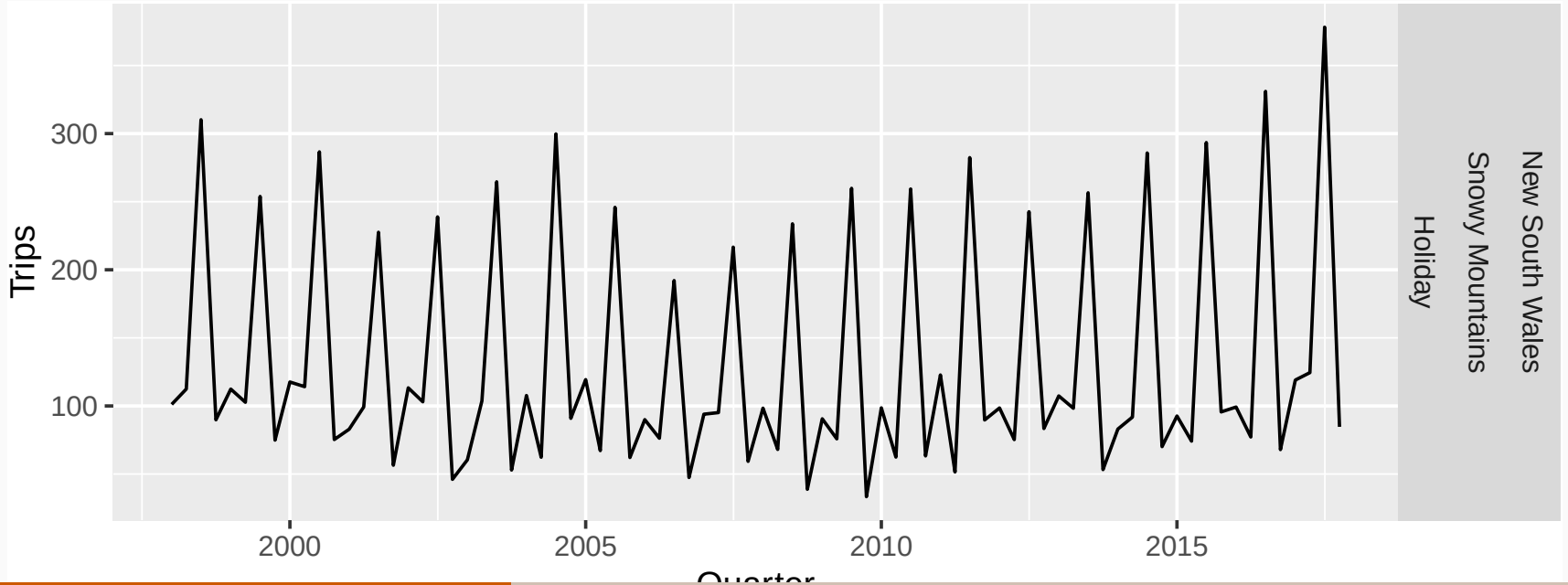
```
most_seasonal <- tourism %>%
  features(Trips, feat_stl) %>%
  filter(seasonal_strength_year == max(seasonal_strength_year))
```

Feature extraction and statistics

Find the most seasonal time series:

```
most_seasonal <- tourism %>%
  features(Trips, feat_stl) %>%
  filter(seasonal_strength_year == max(seasonal_strength_year))
```

```
tourism %>%
  right_join(most_seasonal, by = c("State", "Region", "Purpose")) %>%
  ggplot(aes(x = Quarter, y = Trips)) + geom_line() +
  facet_grid(vars(State, Region, Purpose))
```



Feature extraction and statistics

Find the most trended time series:

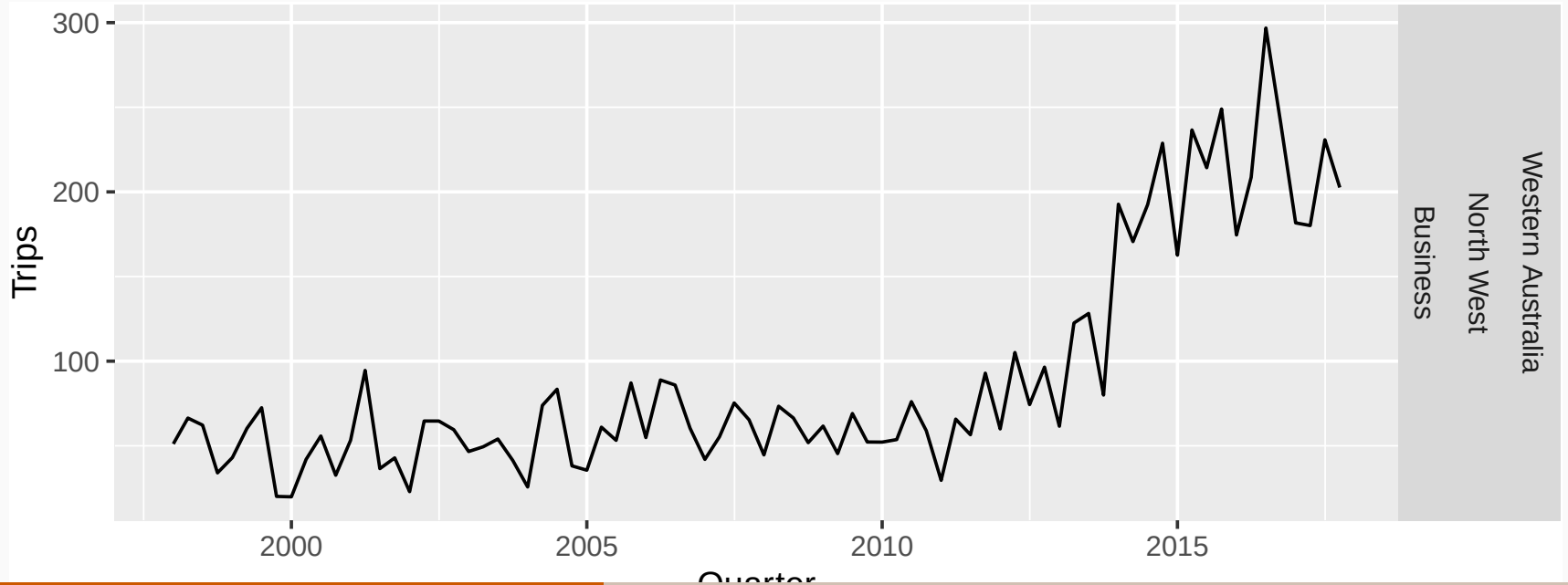
```
most_trended <- tourism %>%
  features(Trips, feat_stl) %>%
  filter(trend_strength == max(trend_strength))
```

Feature extraction and statistics

Find the most trended time series:

```
most_trended <- tourism %>%
  features(Trips, feat_stl) %>%
  filter(trend_strength == max(trend_strength))
```

```
tourism %>%
  right_join(most_trended, by = c("State", "Region", "Purpose")) %>%
  ggplot(aes(x = Quarter, y = Trips)) + geom_line() +
  facet_grid(vars(State, Region, Purpose))
```



Outline

Lab Session 9

- Use `GGally::ggpairs()` to look at the relationships between the STL-based features. You might wish to change `seasonal_peak_year` and `seasonal_trough_year` to factors.
- Which is the peak quarter for holidays in each state?

Feature extraction and statistics

```
tourism %>% features(Trips, feat_acf)
```

```
## # A tibble: 304 x 10
##    Region State Purpose     acf1 acf10 diff1_acf1
##    <chr>  <chr> <chr>      <dbl> <dbl>      <dbl>
##  1 Adela~ Sout~ Busine~  0.0333  0.131     -0.520
##  2 Adela~ Sout~ Holiday  0.0456  0.372     -0.343
##  3 Adela~ Sout~ Other    0.517   1.15      -0.409
##  4 Adela~ Sout~ Visiti~  0.0684  0.294     -0.394
##  5 Adela~ Sout~ Busine~  0.0709  0.134     -0.580
##  6 Adela~ Sout~ Holiday  0.131   0.313     -0.536
##  7 Adela~ Sout~ Other    0.261   0.330     -0.253
##  8 Adela~ Sout~ Visiti~  0.139   0.117     -0.472
##  9 Alice~ Nort~ Busine~  0.217   0.367     -0.500
## 10 Alice~ Nort~ Holiday -0.00660 2.11      -0.153
## # ... with 294 more rows, and 4 more variables:
## #   diff1_acf10 <dbl>, diff2_acf1 <dbl>, diff2_acf10 <dbl>,
## #   season_acf1 <dbl>
```

Outline

Feature extraction and statistics

```
tourism_features <- tourism %>%
  features(Trips, feature_set(pkgs = "feasts"))
```

All features from the feasts package

```
## # A tibble: 304 x 47
##    Region State Purpose trend_strength seasonal_streng~
##    <chr>  <chr> <chr>             <dbl>            <dbl>
##  1 Adela~ Sout~ Busine~           0.451            0.380
##  2 Adela~ Sout~ Holiday           0.541            0.601
##  3 Adela~ Sout~ Other             0.743            0.189
##  4 Adela~ Sout~ Visiti~           0.433            0.446
##  5 Adela~ Sout~ Busine~           0.453            0.140
##  6 Adela~ Sout~ Holiday           0.512            0.244
##  7 Adela~ Sout~ Other             0.584            0.374
##  8 Adela~ Sout~ Visiti~           0.481            0.228
##  9 Alice~ Nort~ Busine~           0.526            0.224
## 10 Alice~ Nort~ Holiday           0.377            0.827
## # ... with 294 more rows, and 42 more variables:
## #   seasonal_peak_year <dbl>, seasonal_trough_year <dbl>,
## #   spikiness <dbl>, linearity <dbl>, curvature <dbl>,
## #   stl_e_acf1 <dbl>, stl_e_acf10 <dbl>, acf1 <dbl>,
## #   acf10 <dbl>, diff1_acf1 <dbl>, diff1_acf10 <dbl>,
```

Feature extraction and statistics

```
pcs <- tourism_features %>%
  select(-State, -Region, -Purpose) %>%
  prcomp(scale = TRUE) %>%
  broom::augment(tourism_features)
```

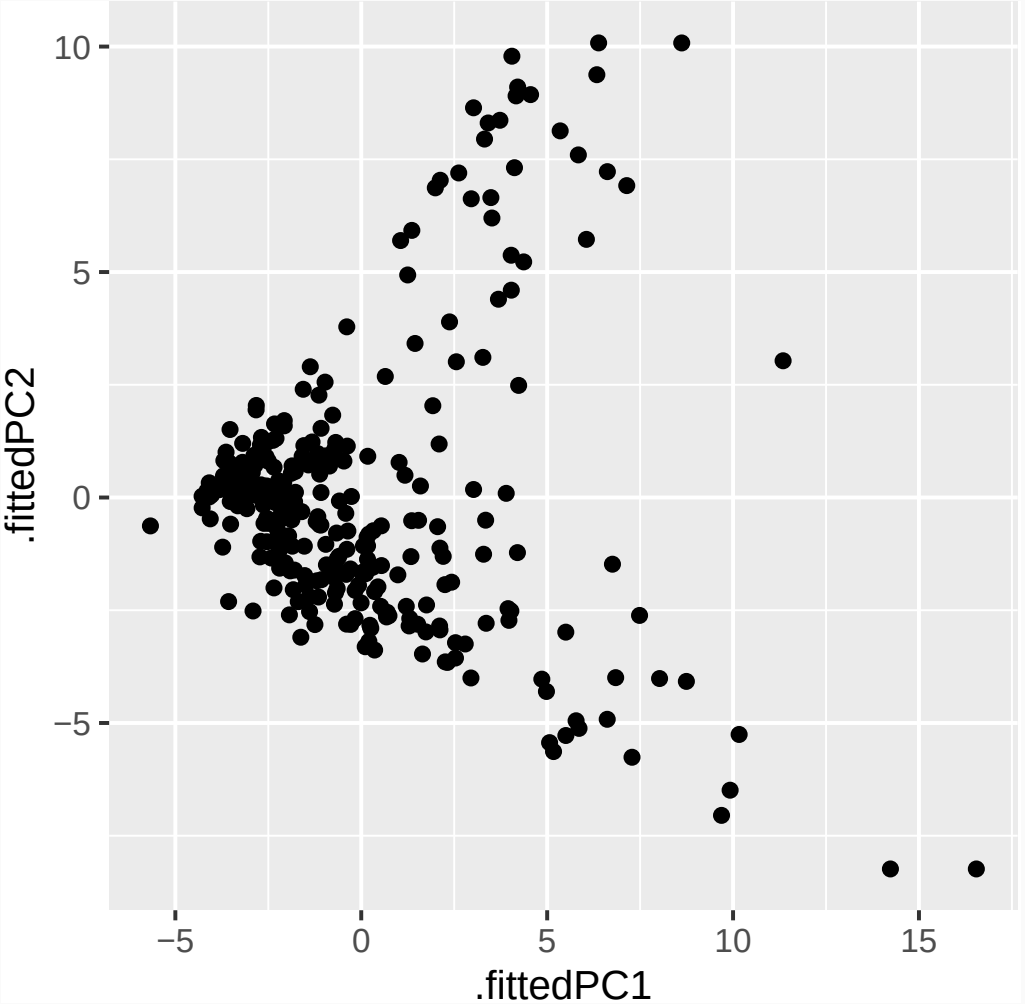
```
## # A tibble: 304 x 92
##    .rownames Region State Purpose trend_strength
##    <fct>     <chr>  <chr> <chr>            <dbl>
##  1 1         Adela~ Sout~ Busine~          0.451
##  2 2         Adela~ Sout~ Holiday          0.541
##  3 3         Adela~ Sout~ Other            0.743
##  4 4         Adela~ Sout~ Visiti~          0.433
##  5 5         Adela~ Sout~ Busine~          0.453
##  6 6         Adela~ Sout~ Holiday          0.512
##  7 7         Adela~ Sout~ Other            0.584
##  8 8         Adela~ Sout~ Visiti~          0.481
##  9 9         Alice~ Nort~ Busine~          0.526
## 10 10        Alice~ Nort~ Holiday          0.377
## # ... with 294 more rows, and 87 more variables:
## #   seasonal_strength_year <dbl>, seasonal_peak_year <dbl>,
## #   seasonal_trough_year <dbl>, spikiness <dbl>,
```

Principal components based on all features from the feasts package

Feature extraction and statistics

```
pcs %>% ggplot(aes(x=.fittedPC1, y=.fittedPC2)) +
  geom_point() + theme(aspect.ratio=1)
```

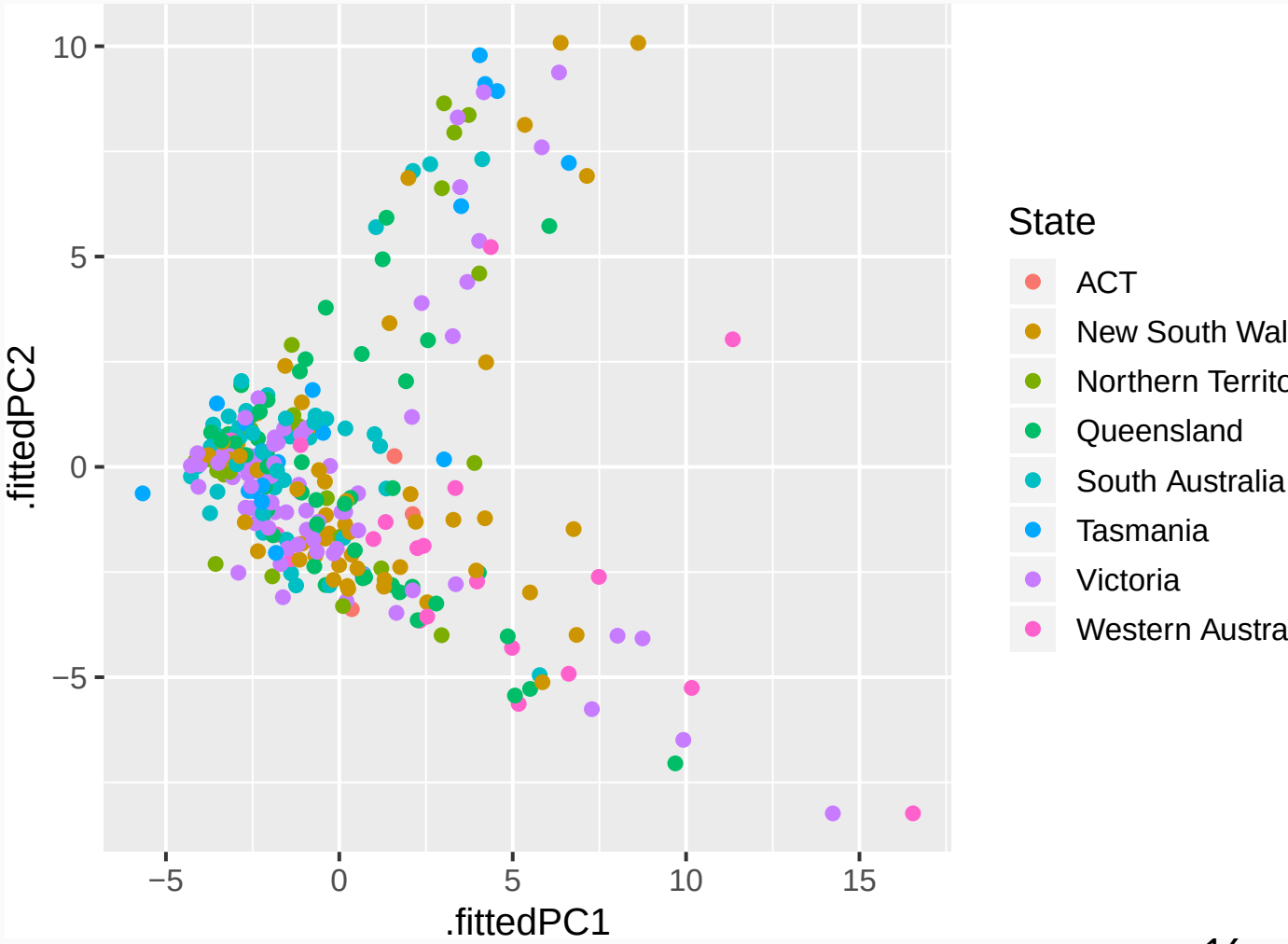
Principal components based on all features from the feasts package



Feature extraction and statistics

```
pcs %>% ggplot(aes(x=.fittedPC1, y=.fittedPC2, col=State)) +
  geom_point() + theme(aspect.ratio=1)
```

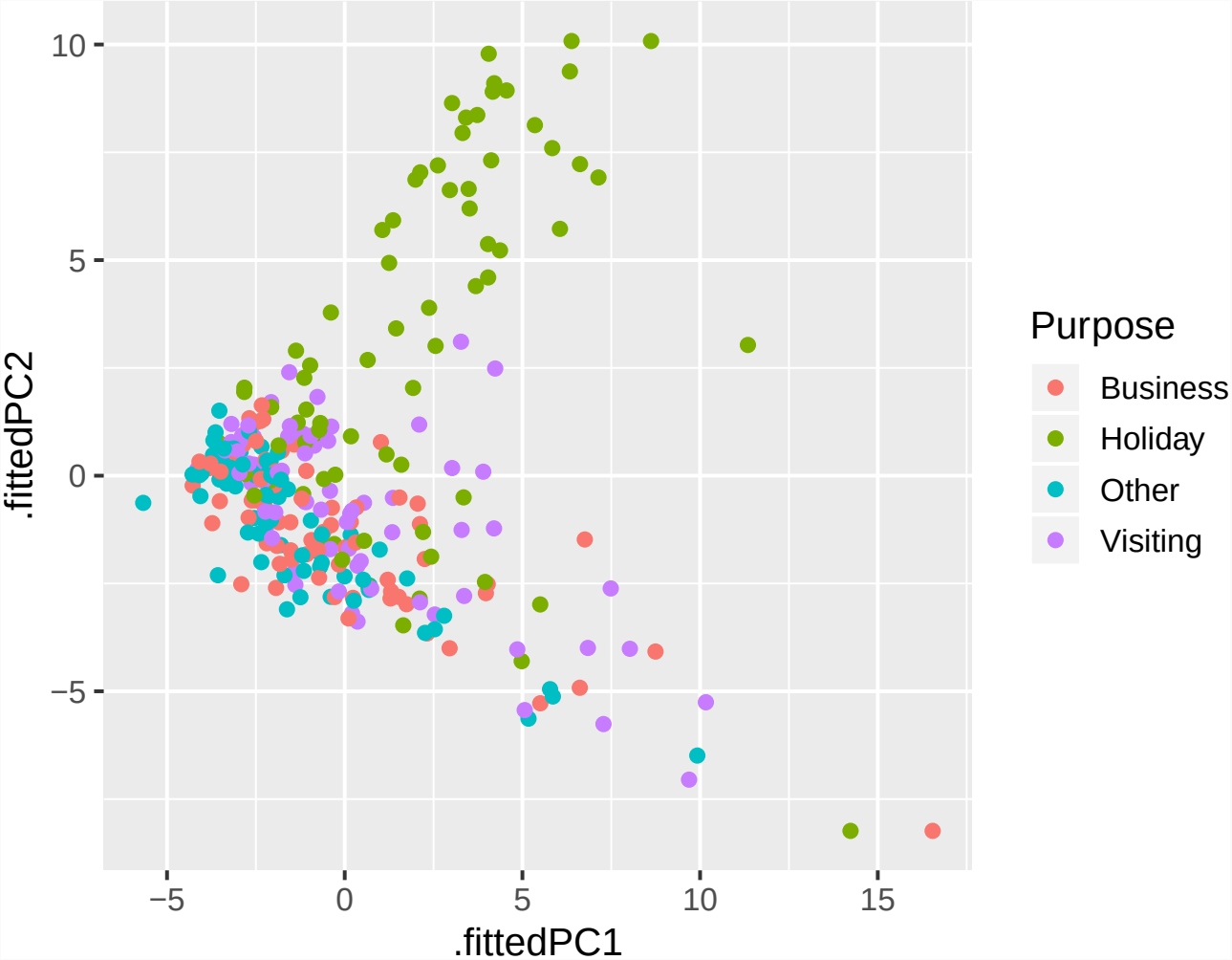
Principal components based on all features from the feasts package



Feature extraction and statistics

```
pcs %>% ggplot(aes(x=.fittedPC1, y=.fittedPC2, col=Purpose)) +
  geom_point() + theme(aspect.ratio=1)
```

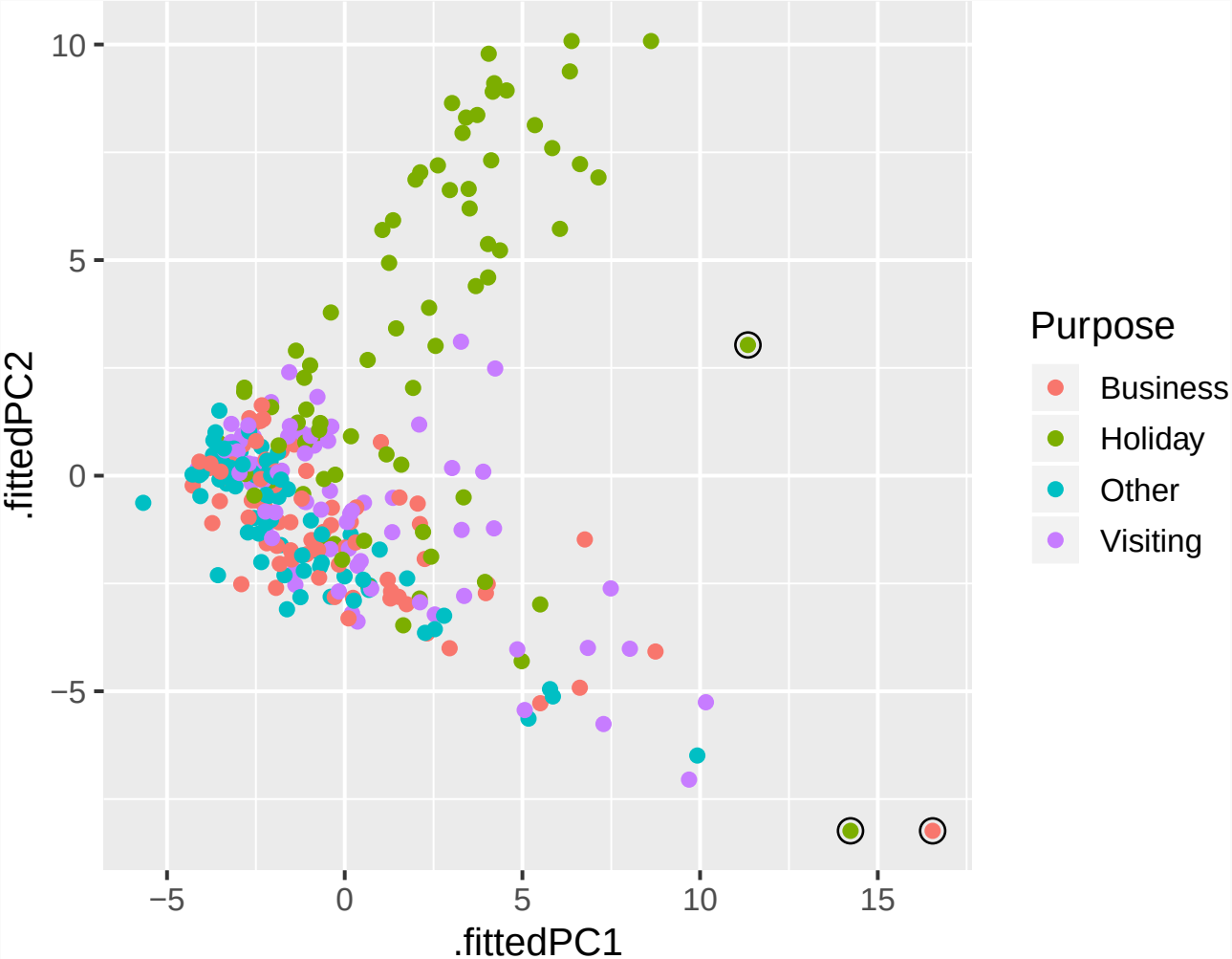
Principal components based on all features from the feasts package



Feature extraction and statistics

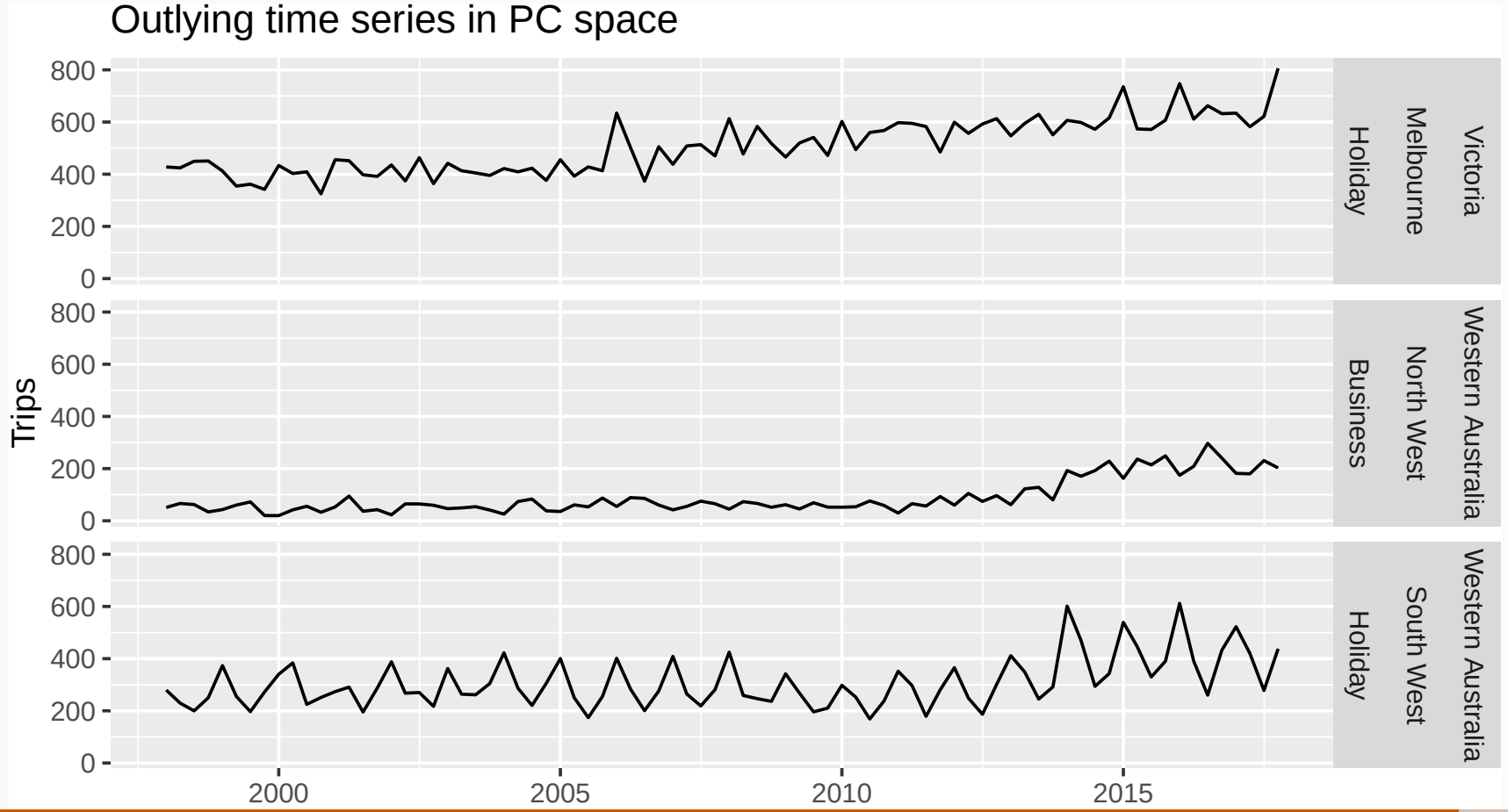
```
pcs %>% ggplot(aes(x=.fittedPC1, y=.fittedPC2, col=Purpose)) +
  geom_point() + theme(aspect.ratio=1)
```

Principal components based on all features from the feasts package



Feature extraction and statistics

```
outliers %>%
  left_join(tourism, by = c("State", "Region", "Purpose")) %>%
  mutate(Series = glue("{State}", "{Region}", "{Purpose}", .sep = "\n\n")) %>%
  ggplot(aes(x = Quarter, y = Trips)) + geom_line() +
  facet_grid(Series ~ .) + ggtitle("Outlying time series in PC space")
```



Outline

Lab Session 10

- Use a feature-based approach to look for outlying series in PBS.
- What is unusual about the series you identify as outliers?