Melanoma Study

Exploratory Analysis & Statistical Inference

What is Melanoma?

- Have melanoma registration rates in New Zealand changed significantly over time, and how do they compare across gender groups and environmental UV exposure periods?

Dataset source and description

- Melanoma registration rates, by age group, 1996–2015, Obtained from data.govt.nz
- Estimated resident population (2023-base): At 30 June 2023, Obtained from stats.govt.nz
- Daily peak UV index value, 1981–2017, Obtained from data.govt.nz

Data Validation

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1996    2001    2006    2006    2010    2015
[1] "All"    "Female" "Male"
[1] "0-24"     "25-44"    "45-64"    "65-74"    "75+"      "All_ages"
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.6333  24.7012  61.2380  85.7679 137.5900 354.7888
```

```r
# Check Year column for missing entries
MissingValuesCheck(melanomaRawDF$Year)

# Check Gender column for missing entries
MissingValuesCheck(melanomaRawDF$Gender)

# Check Age_group column for missing entries
MissingValuesCheck(melanomaRawDF$Age_group)

# Check Registration_rate column for missing entries
MissingValuesCheck(melanomaRawDF$Registration_rate)
```

```
[1] 0
[1] 0
[1] 0
[1] 0
```

Data Validation

```
# Check Location column for unique entries, ensure only expected values exist
unique(uvIndexDF$Location)

# Check Daily_Peak_UVI column for range of values (make sure no garbage outliers)
summary(uvIndexDF$Daily_peak_UVI)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    8.00   16.00   15.74   23.00   31.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   7.000   6.564  10.000  12.000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1981    2000    2006    2005    2011    2017
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    94.0   186.0   184.4   277.0   366.0
[1] "Christchurch" "Invercargill" "Lauder"       "Leigh"        "Paraparaumu"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.1     1.6     3.9     5.0     8.1    16.8
```

```{r}
# Check Date column for missing entries
MissingValuesCheck(uvIndexDF$Date)
```

Data Validation

	A	B	C	D	E	F	G	H	I	J	K
1	Table 1										
3	**Estimated population by sex**										
4	1926–2024										
5	Year ended 31 December										
7	Year	Population at 31 December			Estimated increase/ decrease during year		Sex ratio at 31 December[1]	Mean population for year ended 31 December			
8		Males	Females	Total	Number	Percent		Males	Females	Total	
9					**Estimated de facto population[2]**						
10	1926	730,500	699,200	1,429,700	28,500	2.03	104.5	722,600	691,200	1,413,700	
11	1927	740,500	709,900	1,450,400	20,700	1.45	104.3	735,000	704,000	1,439,000	
12	1928	748,600	718,700	1,467,400	17,000	1.17	104.2	743,000	713,000	1,456,100	
13	1929	757,800	728,300	1,486,100	18,700	1.27	104.1	751,300	722,100	1,473,400	
14	1930	767,900	738,900	1,506,800	20,700	1.39	103.9	761,100	732,000	1,493,000	
15	1931	775,600	747,100	1,522,800	16,000	1.06	103.8	771,700	742,500	1,514,200	
16	1932	780,900	753,800	1,534,700	11,900	0.78	103.6	777,500	749,600	1,527,100	
17	1933	786,400	760,800	1,547,100	12,400	0.81	103.4	783,000	756,600	1,539,600	
18	1934	792,000	766,400	1,558,400	11,300	0.73	103.3	788,600	762,900	1,551,500	
19	1935	796,700	773,000	1,569,700	11,300	0.73	103.1	793,600	768,700	1,562,200	
20	1936	804,300	780,300	1,584,600	14,900	0.95	103.1	799,700	775,500	1,575,200	
21	1937	813,100	788,700	1,601,800	17,200	1.09	103.1	807,300	782,700	1,590,000	
22	1938	821,700	796,600	1,618,300	16,500	1.03	103.2	815,700	791,000	1,606,800	
23	1939	832,800	808,800	1,641,600	23,300	1.44	103.0	826,800	801,700	1,628,500	

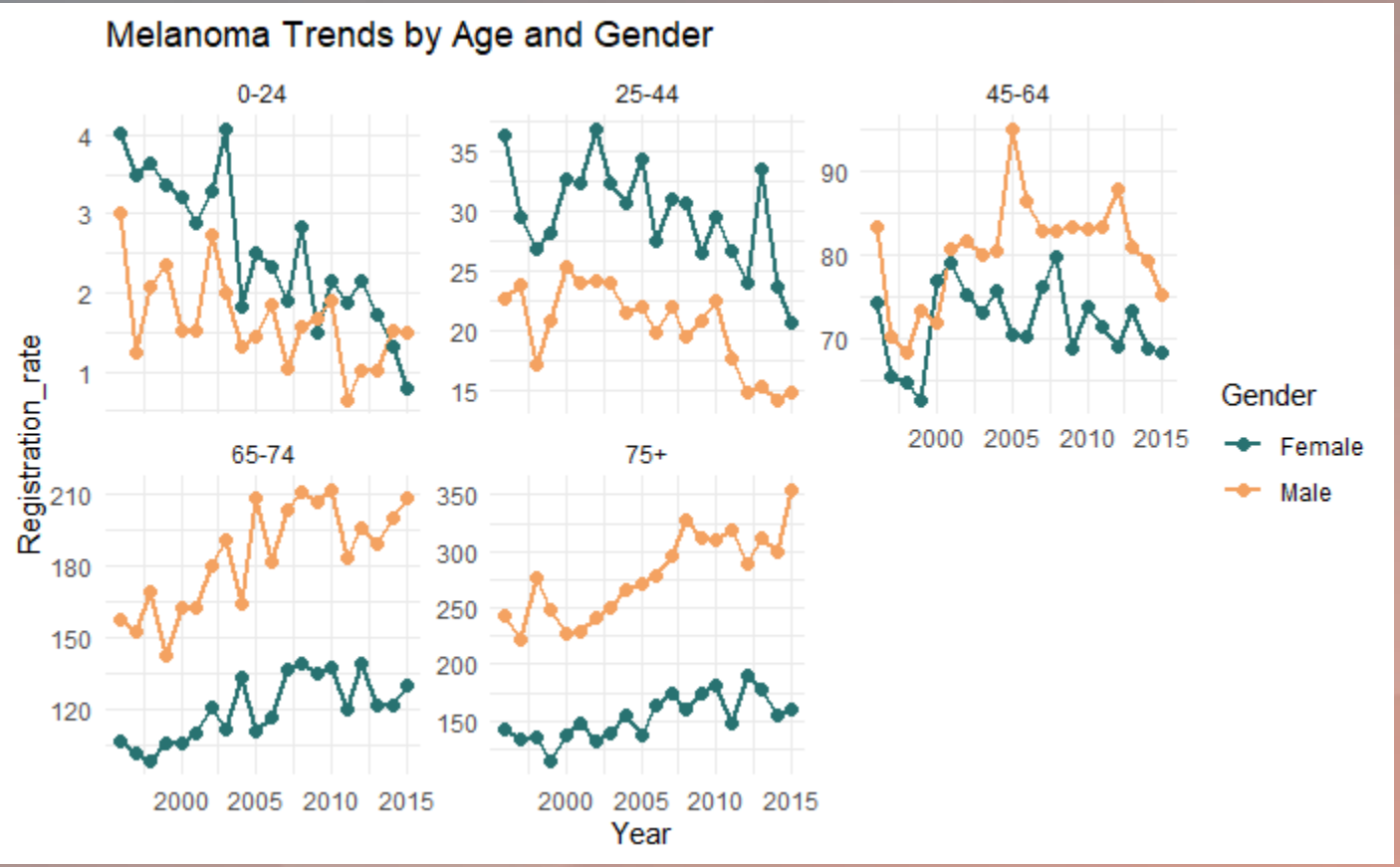
Data Validation

```r
# Check mean_male_pop column for unique entries, ensure only expected values exist
summary(populationRawDF$mean_male_pop)

# Check mean_female_pop column for unique entries, ensure only expected values exist
summary(populationRawDF$mean_female_pop)

# Check mean_total_pop column for range of values (make sure no garbage outliers)
summary(populationRawDF$mean_total_pop)
```

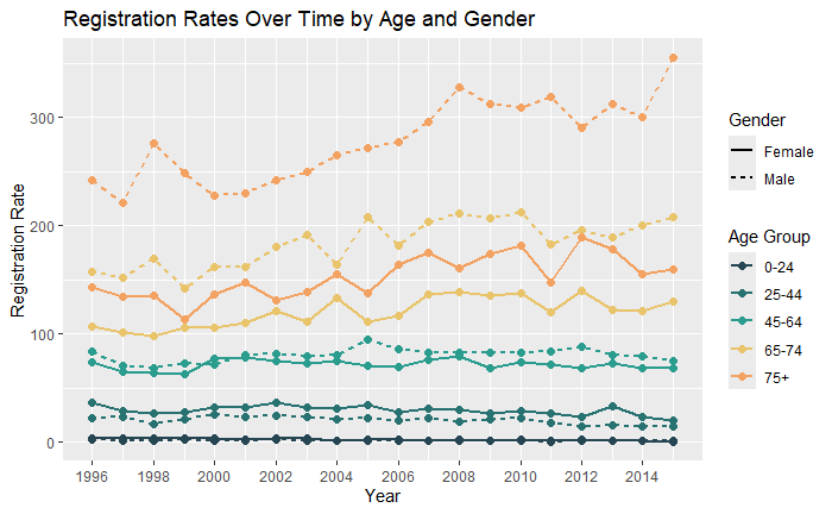
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1926    1950    1975    1975    2000    2024 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 730500  978400 1567600 1504707 1896050 2642400 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 699200  970700 1576100 1527389 1966050 2668700 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
1429700 1949100 3143700 3032097 3862100 5311100 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -8000   20700   38200   38843   51500  142700 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -0.490   0.810   1.210   1.338   1.945   3.090 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  92.90   97.10   99.10   99.15  100.95  104.50 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 722600  968100 1547600 1492107 1890500 2633500 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 691200  960250 1552500 1514487 1958200 2655600 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
1413700 1928350 3100100 3006599 3848750 5289100 
```

```r
# Check year column for missing entries
MissingValuesCheck(populationRawDF$year)
```

Melanoma Analysis

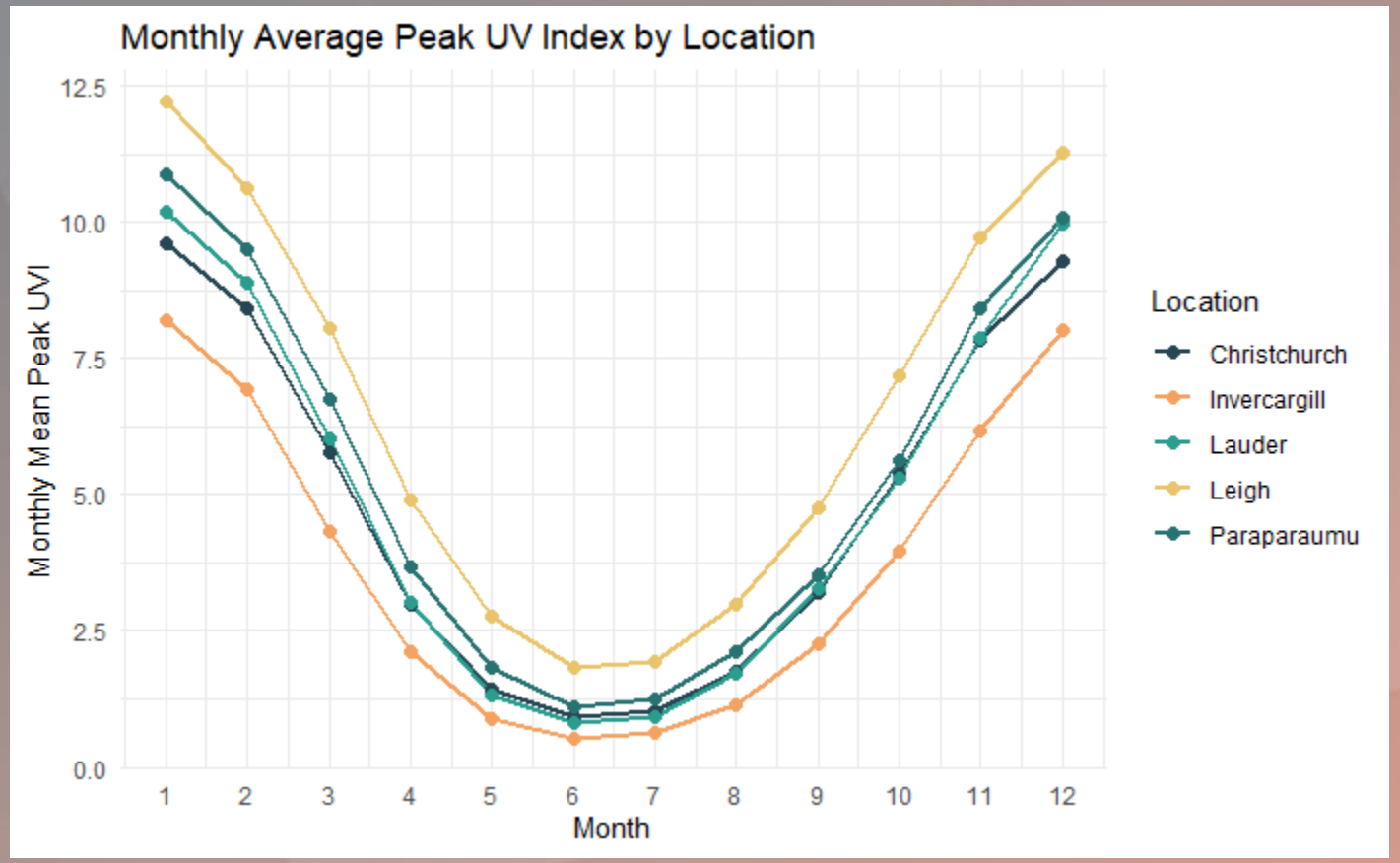


Melanoma Analysis

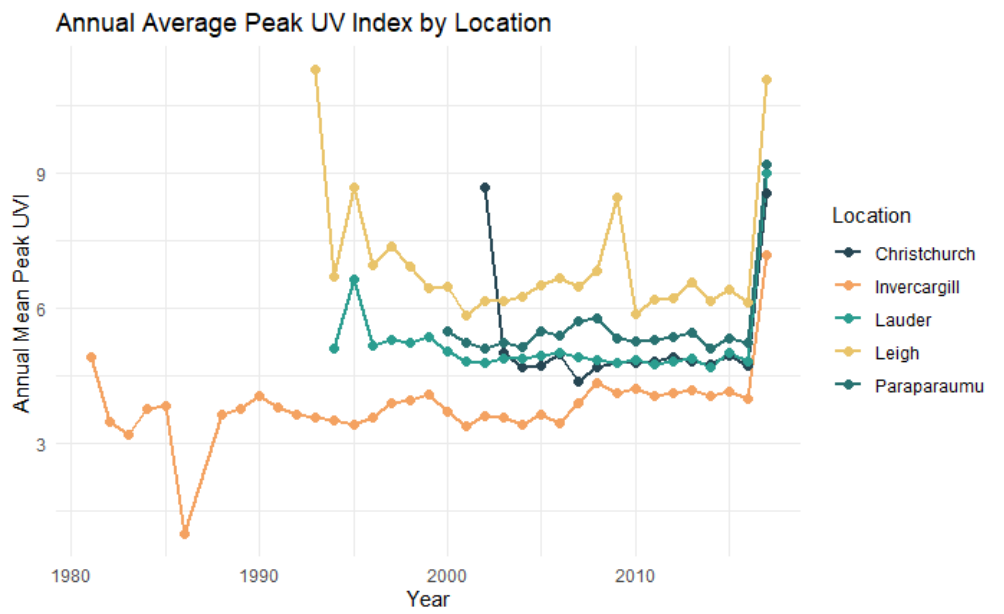


Age_group <chr>	Gender <chr>	mean_rate <dbl>	sd_rate <dbl>
0-24	Female	2.541044	0.9351106
0-24	Male	1.647825	0.5872297
25-44	Female	29.659574	4.2327945
25-44	Male	20.349196	3.5259057
45-64	Female	71.816306	4.6958514
45-64	Male	80.509676	6.3228605
65-74	Female	120.188186	13.4920452
65-74	Male	184.053649	21.7982217
75+	Female	152.940278	19.8236945
75+	Male	278.515234	37.6321607

UV Index Analysis

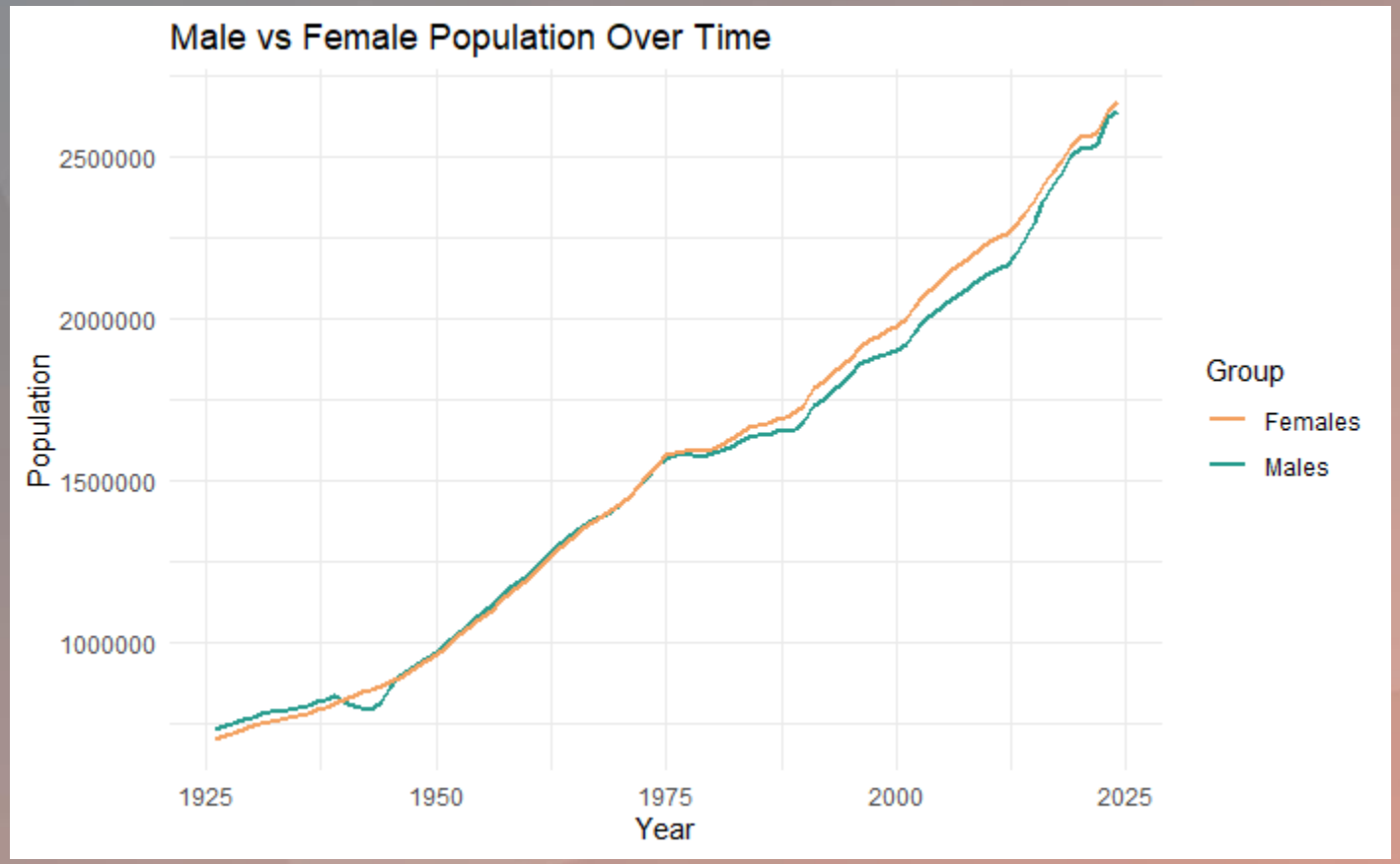


UV Index Analysis

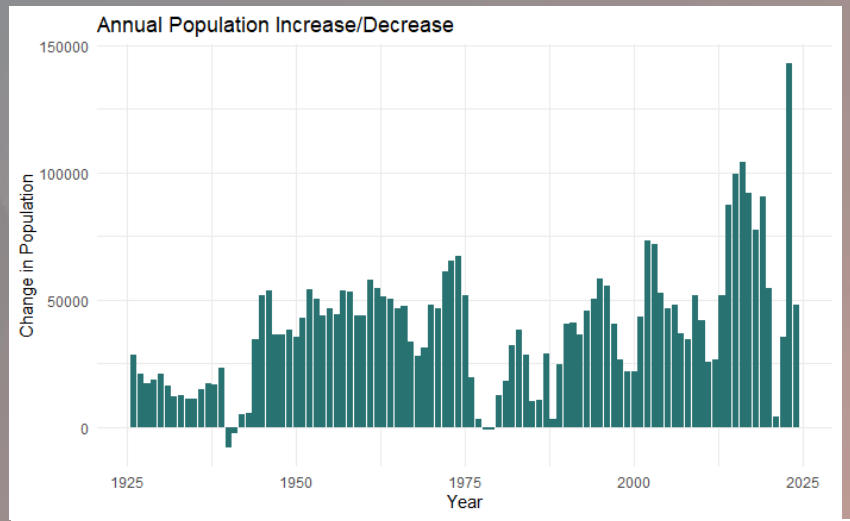


Location <chr>	1980–1989 <dbl>	1990–1999 <dbl>	2000–2009 <dbl>	2010–2019 <dbl>
Christchurch	*NA*	*NA*	4.833786	4.900191
Invercargill	3.53707	3.731692	3.713917	4.171549
Lauder	*NA*	5.392932	4.885390	4.916360
Leigh	*NA*	7.088632	6.524484	6.337446
Paraparaumu	*NA*	*NA*	5.389066	5.379268

Population Analysis



Population Analysis

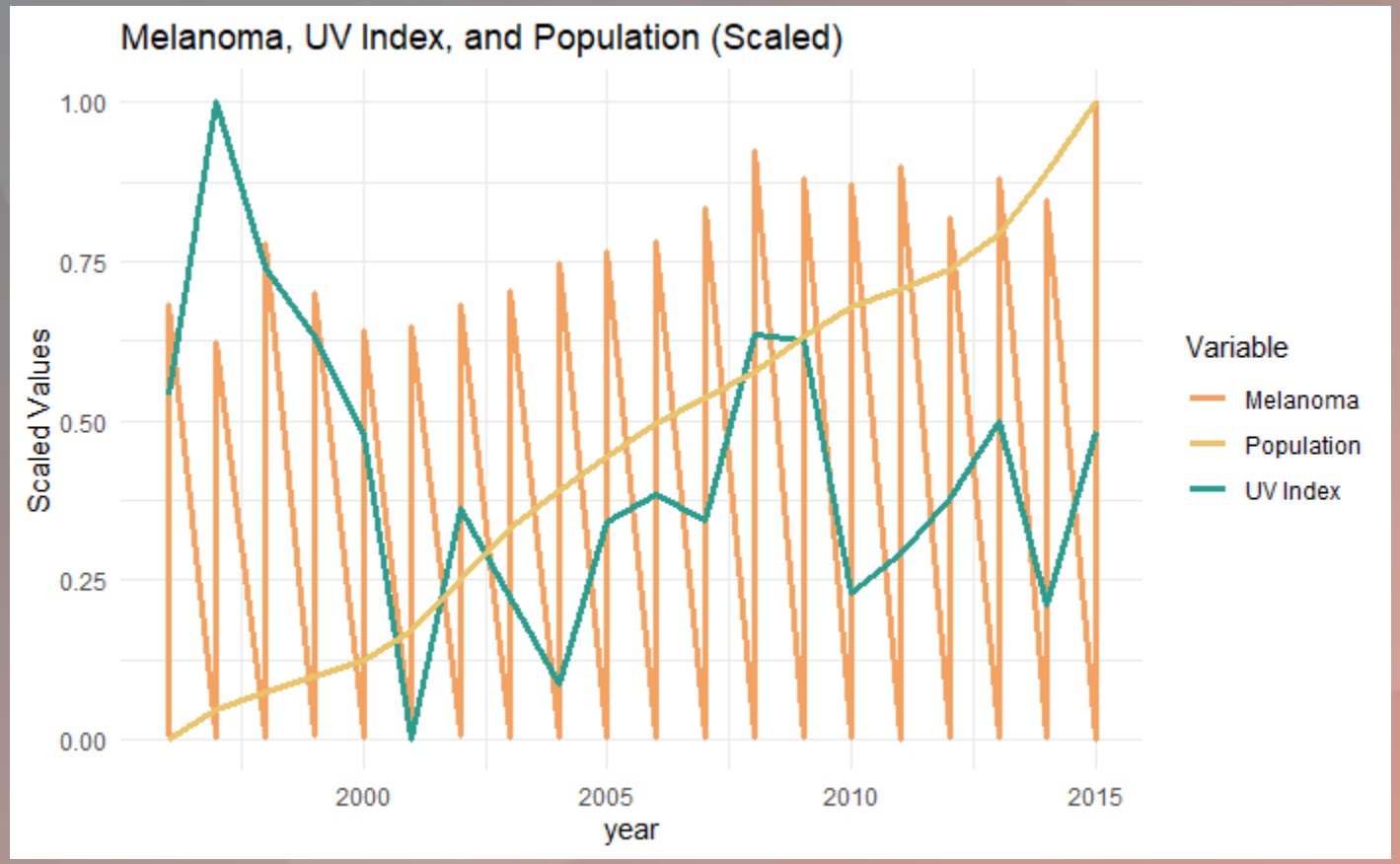


decade <chr>	mean_pop <dbl>	mean_male <dbl>	mean_female <dbl>
1920–1929	1458400	744350	714025
1930–1939	1568580	797140	771440
1940–1949	1729200	854600	874600
1950–1959	2142840	1077290	1065550
1960–1969	2626170	1316720	1309450
1970–1979	3062970	1527830	1535140
1980–1989	3282920	1626150	1656770
1990–1999	3667670	1806260	1861410
2000–2009	4118290	2017220	2101090
2010–2019	4649560	2290060	2359490

Merged Data

```r
450  ```{r}
451  # Filtered UVI dataframe, from within time period,
452  filteredUVI <- uvIndexDF %>%
453    filter(year >= 1996 & year <= 2015) %>%
454    group_by(year) %>%
455    summarise(avg_uvi = mean(Daily_peak_UVI))
456
457  # Filtered Population Dataframe, from within time period
458  filteredPopulation <- populationDF %>%
459    filter(year >= 1996 & year <= 2015)
460
461  # Filtered Melanoma Dataframe; from within time period
462  filteredMelanoma <- melanomaDF %>%
463    rename(year = Year) %>%
464    rename(gender = Gender) %>%
465    rename(age_group = Age_group) %>%
466    rename(registration_rate = Registration_rate) %>%
467    filter(year >= 1996 & year <= 2015)
468
469  # Merge all three dataframes
470  mergedDF <- filteredMelanoma %>%
471    left_join(filteredUVI, by = "year") %>%
472    left_join(filteredPopulation, by = "year")
473
```

Merged Analysis



Hypothesis Testing

- **H_0 (Null):** There is *no difference* in mean melanoma registration rates between the early period (1996–2005) and the late period (2006–2015).

- **H_0 (Null) :** Mean melanoma rates are *equal* for males and females.

- **H_0 (Null)::** Melanoma rates do *not differ* between high-UV and low-UV years.

Hypothesis Testing

- **H_0 (Null):** There is *no difference* in mean melanoma registration rates between the early period (1996–2005) and the late period (2006–2015).
- **H_1 (Alt):** There *is a difference* in mean melanoma registration rates between the early period (1996–2005) and the late period (2006–2015).

```
        Welch Two Sample t-test

data:  earlySample and lateSample
t = -1.1049, df = 188.18, p-value = 0.2706
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -37.98683  10.71021
sample estimates:
mean of x mean of y
 87.40294 101.04125
```

Hypothesis Testing

- **H_0:** Mean melanoma rates are *equal* for males and females.
- **H_1:** Mean melanoma rates are *different* between males and females.

```
        Welch Two Sample t-test

data:  male_rates and female_rates
t = 3.109, df = 151.09, p-value = 0.002244
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 13.69978 61.47229
sample estimates:
mean of x mean of y
113.01512  75.42908
```

Hypothesis Testing

- **H_0:** Melanoma rates do *not differ* between high-UV and low-UV years.
- **H_1:** Melanoma rates are *different* between high-UV and low-UV years.

```
        Welch Two Sample t-test

data:  melanoma_rate by uv_group
t = -0.88527, df = 14.618, p-value = 0.3903
alternative hypothesis: true difference in means between group High UV and group Low UV is not equal to 0
95 percent confidence interval:
 -11.148848   4.616032
sample estimates:
mean in group High UV  mean in group Low UV
             92.58889              95.85530
```

Future Analysis

- Wider time periods required due to latency factors
- Additional information about patients needed (sun exposure amount, age of sun exposure etc)
- Ethnicity and Gender

References

- https://catalogue.data.govt.nz/dataset/melanoma-registration-rates-by-age-group-19962015
- https://www.stats.govt.nz/information-releases/estimated-resident-population-2023-base-at-30-june-2023/
- https://catalogue.data.govt.nz/dataset/daily-peak-uv-index-value-19812017
- https://niwa.co.nz/atmosphere/uv-and-ozone/uv-index-information