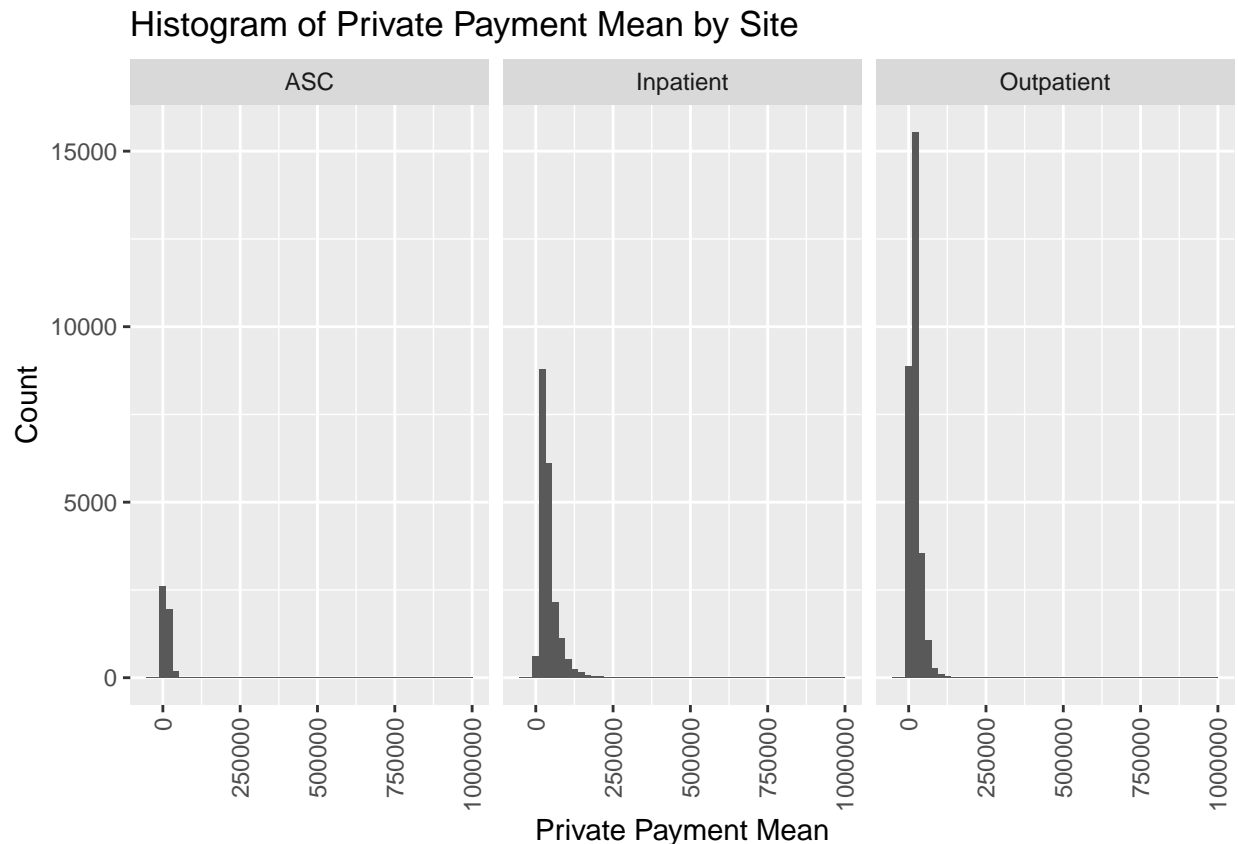


# Capstone EDA

## Exploratory Data Analysis

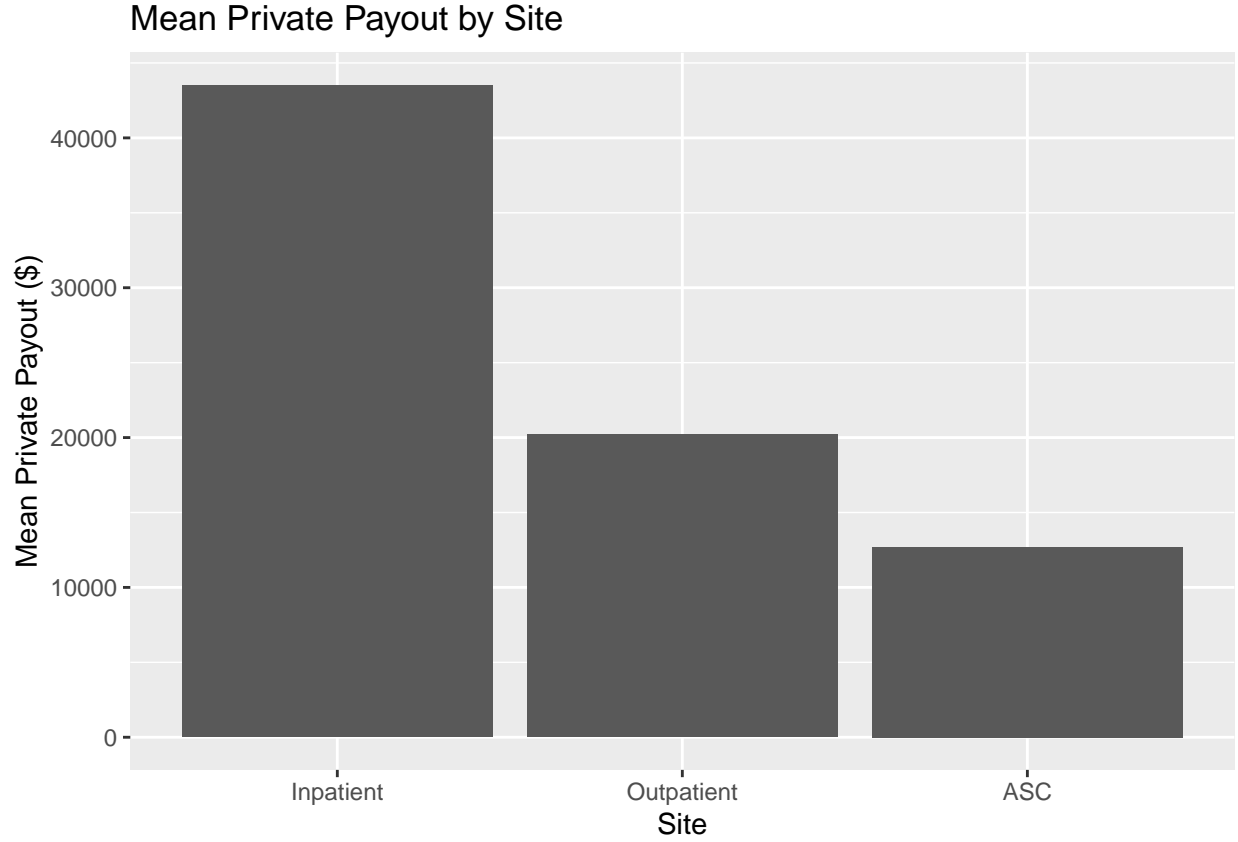
To start, we wanted to look at some of the features that were helpful in predicting `priv_pay_mean` in the original model done by JnJ. First, we wanted to see how `priv_pay_mean` varied by `site`. To do this, we created a histogram faceted by site as well as a table with summary statistics by site.



From the histogram, we can see that there is clearly a right-skew in the data regardless of site. This is unsurprising since some procedures can be very expensive which would result in higher `priv_pay_mean`, but there isn't going to be symmetry since there is not any procedure where `priv_pay_mean` will be very negative. These histograms also show us that ASC is the least common site used for surgery, and outpatient surgeries are more common than inpatient in our data. From the distribution of the three histograms, it looks like inpatient surgeries typically have the highest `priv_pay_mean`, followed by outpatient, then ASC being the cheapest. This is what we expected since inpatient surgeries involve the patient staying overnight which leads to `priv_pay_mean` being higher typically. It also makes sense that outpatient `priv_pay_mean` would be higher than ASC `priv_pay_mean` since outpatient surgeries happen at the hospital whereas ASC procedures are at a surgery center that is not a hospital. Outpatient surgery being part of a hospital-run facility leads to higher `priv_pay_mean` usually which is what we see in the data provided by Johnson and Johnson. The table reiterates some of the points mentioned above, and also looks at medicare payments by site. The same trends we saw with `priv_pay_mean` also exist with medicare payments in our data.

Table 1: Summary Statistics by Site

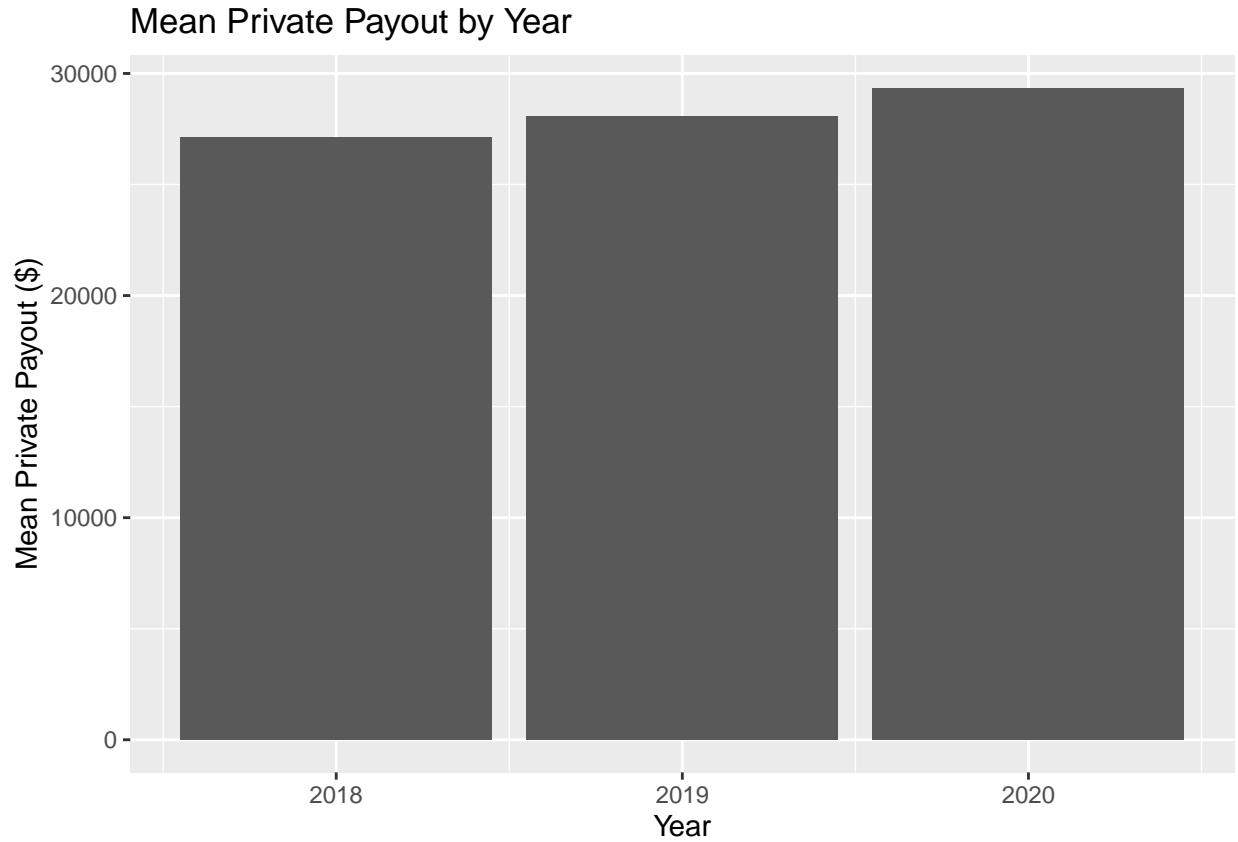
| site       | site_mean_priv_pay | site_median_priv_pay | site_mean_mcare_pay | site_median_mcare_pay | total_priv_count |
|------------|--------------------|----------------------|---------------------|-----------------------|------------------|
| Inpatient  | 43509.48           | 32910.476            | 20594.748           | 19342.245             | 212180           |
| Outpatient | 20213.08           | 15244.762            | 8124.497            | 8196.171              | 600142           |
| ASC        | 12693.11           | 8920.038             | 5943.262            | 5678.334              | 113085           |



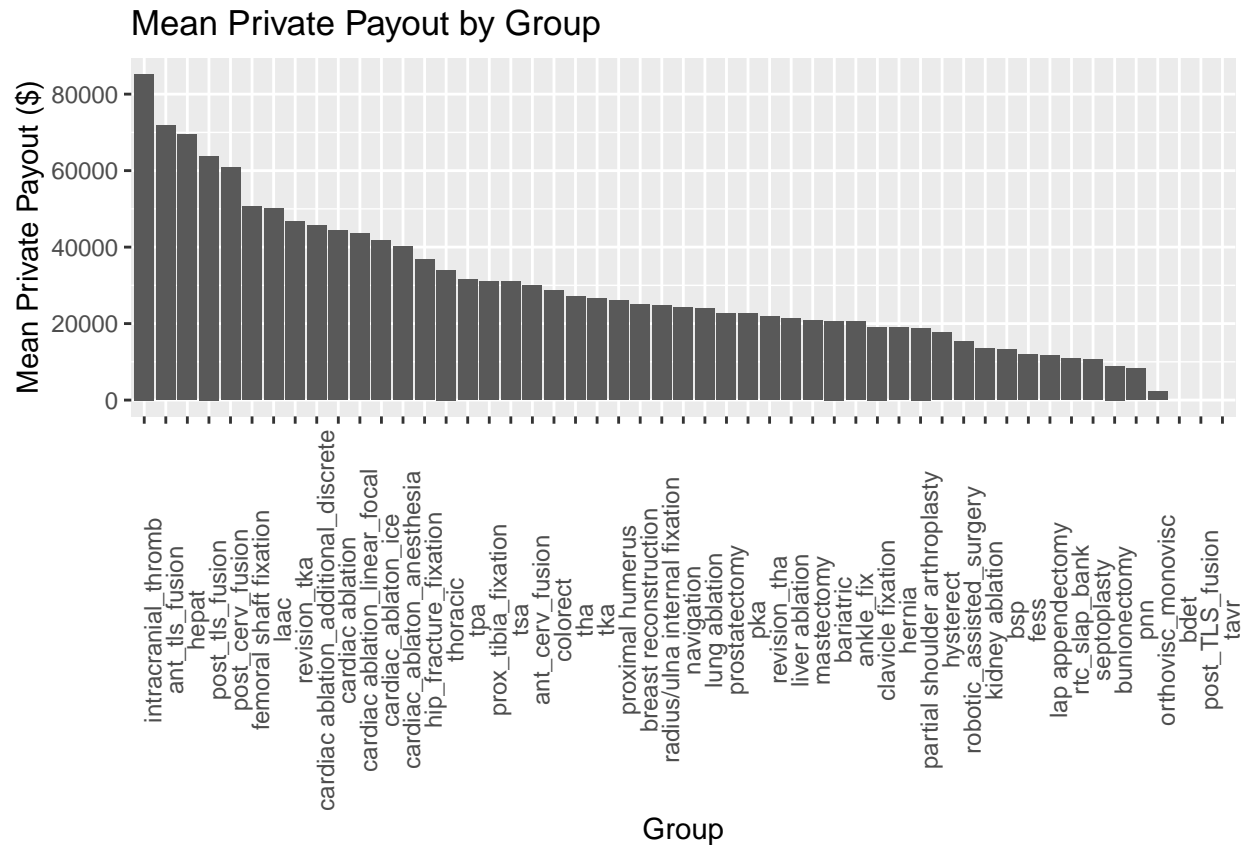
Next, we looked at variation in payment by year since the data included 2018-2020. We did not expect much to change from year-to-year, but did expect payments to increase slightly due to inflation. From the table, both private and medicare payments have increased slightly each year from 2018 to 2020. Another interesting thing to note is that `priv_count` decreased each year in the data. It seems logical that the number of procedures would have decreased in 2020 due to COVID, but it is not as clear why there was also a decrease from 2018 to 2019.

Table 2: Summary Statistics by Year

| year | year_mean_priv_pay | year_median_priv_pay | year_mean_mcare_pay | year_median_mcare_pay | total_priv_count |
|------|--------------------|----------------------|---------------------|-----------------------|------------------|
| 2018 | 27121.98           | 20108.73             | 7422.005            | 7124.566              | 372001           |
| 2019 | 28065.65           | 20778.76             | 8034.075            | 7624.272              | 303315           |
| 2020 | 29353.92           | 21597.70             | 9009.511            | 8662.794              | 250091           |



We also wanted to look at how payment varied by group since different surgeries vary in cost, and thus also vary in payment. Looking at the table below, there are 51 different types of surgery in the data with `rtc_slap_bank` being the most common based on `priv_count`. There is large variation in average private payment by group with the lowest averaging just over \$2200 and the highest averaging just over \$85000.



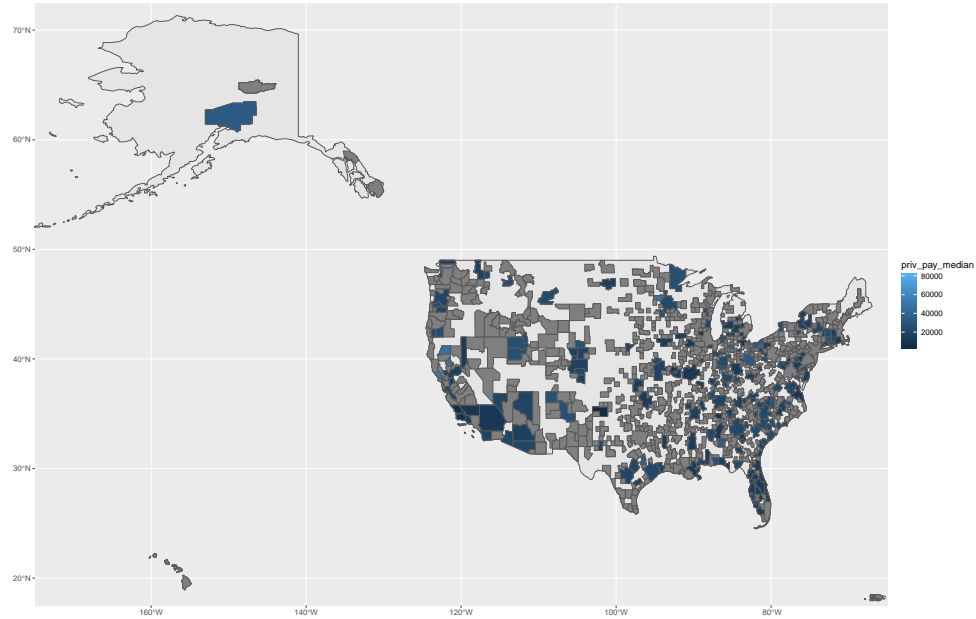
## Looking at geographic regions

Does geography seem to have an impact on insurance payouts? Let's take a look. First, we will look at the average private payment for tka in 2018, a group and year that is quite well-represented.

```
msa_geo <- read_sf("./shapefiles_cbsa", "cb_2021_us_cbsa_20m")
us_geo <- read_sf("./shapefiles_nation", "cb_2021_us_nation_20m")
msa_geo$GEOID = as.integer(msa_geo$GEOID)
tka_2020_data <- data %>% filter(group == "tka" & year == 2020)
tka_2020_data <- full_join(tka_2020_data, msa_geo, by=c("msa" = "GEOID"))
ggplot(data=tka_2020_data) +
  geom_sf(data=us_geo, aes(geometry=geometry)) +
  geom_sf(aes(fill=priv_pay_median, geometry=geometry)) +
  coord_sf(xlim=c(-170,-70), ylim = c(20,70))
```

Table 3: Summary Statistics by Group

| group                                | group_mean_priv_pay | group_median_priv_pay | group_mean_mcare_pay | group_median_mcare_pay | total_priv_count |
|--------------------------------------|---------------------|-----------------------|----------------------|------------------------|------------------|
| intracranial_thromb                  | 85256.312           | 72161.250             | 28498.726            | 29088.644              | 796              |
| ant_tls_fusion                       | 71864.892           | 64185.967             | 33526.048            | 31884.588              | 3909             |
| hepat                                | 69549.109           | 48815.951             | 19438.154            | 13807.790              | 6410             |
| post_tls_fusion                      | 63839.024           | 62752.949             | 28729.442            | 27507.392              | 20888            |
| post_cerv_fusion                     | 60761.972           | 50711.060             | 25513.193            | 22493.884              | 2730             |
| femoral_shaft_fixation               | 50588.634           | 37022.795             | 17719.976            | 16675.590              | 1567             |
| laac                                 | 50207.853           | 44782.530             | 16317.891            | 16650.441              | 146              |
| revision_tka                         | 46656.752           | 40946.265             | 22408.088            | 21669.410              | 98               |
| cardiac_ablation_additional_discrete | 45574.914           | 41532.000             | 21163.217            | 21450.404              | 5333             |
| cardiac_ablation                     | 44266.388           | 40399.310             | 23746.398            | 23782.747              | 23744            |
| cardiac_ablation_linear_focal        | 43500.701           | 40288.448             | 21607.005            | 22109.825              | 3728             |
| cardiac_ablation_ice                 | 41646.821           | 39375.380             | 21547.004            | 21933.067              | 14038            |
| cardiac_ablation_anesthesia          | 40226.110           | 37768.905             | NaN                  | NaN                    | 15332            |
| hip_fracture_fixation                | 36859.235           | 30885.920             | 14561.084            | 13380.670              | 2445             |
| thoracic                             | 34022.736           | 30932.121             | 16826.731            | 15821.986              | 2320             |
| tpa                                  | 31594.245           | 28119.810             | 15532.311            | 15108.532              | 2935             |
| prox_tibia_fixation                  | 31096.575           | 23470.170             | 13691.182            | 13046.771              | 4535             |
| tta                                  | 31091.793           | 29814.190             | 13306.561            | 13410.252              | 4826             |
| ant_cerv_fusion                      | 29904.385           | 26432.139             | 8480.087             | 8406.347               | 20451            |
| colorect                             | 28721.916           | 26574.884             | 17147.491            | 13830.769              | 21918            |
| tha                                  | 27134.456           | 25793.878             | 14221.365            | 13652.864              | 39153            |
| tka                                  | 26614.780           | 24857.907             | 9446.357             | 9504.127               | 55269            |
| proximal_humerus                     | 25954.878           | 20024.944             | 14294.652            | 14186.412              | 3685             |
| breast_reconstruction                | 25001.692           | 20152.958             | 9293.869             | 9088.802               | 44591            |
| radius/ulna_internal_fixation        | 24837.718           | 15228.101             | 10828.827            | 10351.535              | 15199            |
| navigation                           | 24247.969           | 22244.785             | 12400.528            | 12507.221              | 2334             |
| lung_ablation                        | 23964.086           | 12037.610             | 10337.743            | 10351.706              | 92               |
| prostatectomy                        | 22747.804           | 21186.280             | 10027.186            | 10165.426              | 6343             |
| pka                                  | 22674.411           | 20295.217             | 7941.219             | 7925.798               | 4896             |
| revision_tha                         | 21874.463           | 14179.720             | 14275.400            | 14193.668              | 964              |
| liver_ablation                       | 21409.303           | 13505.900             | 11073.294            | 10855.424              | 597              |
| mastectomy                           | 20712.359           | 16999.277             | 6491.671             | 6674.743               | 37944            |
| bariatric                            | 20662.184           | 20422.877             | 16733.906            | 15336.532              | 41239            |
| ankle_fix                            | 20643.949           | 13717.686             | 5297.749             | 4001.100               | 26522            |
| clavicle_fixation                    | 19100.965           | 13152.759             | 8634.423             | 8531.798               | 3352             |
| hernia                               | 19041.905           | 13463.575             | 10425.533            | 9139.392               | 24508            |
| partial_shoulder_arthroplasty        | 18856.347           | 13005.200             | 12063.691            | 12391.246              | 1535             |
| hysterect                            | 17714.882           | 15970.557             | 9144.329             | 9294.473               | 90028            |
| robotic_assisted_surgery             | 15385.740           | 13266.420             | 5304.823             | 5211.536               | 21961            |
| kidney_ablation                      | 13575.053           | 11565.220             | 7563.422             | 7323.914               | 241              |
| bsp                                  | 13202.974           | 11962.593             | 4886.031             | 4844.942               | 2399             |
| fess                                 | 11938.086           | 10989.062             | 4223.115             | 4372.276               | 84719            |
| lap_appendectomy                     | 11710.227           | 10834.927             | 4454.832             | 4459.976               | 34408            |
| rtc_slap_bank                        | 10788.735           | 9346.225              | 5568.469             | 5606.205               | 173520           |
| septoplasty                          | 10659.454           | 9721.611              | 3716.994             | 3829.554               | 24149            |
| bunionectomy                         | 8894.394            | 7057.646              | 2405.460             | 2362.119               | 25376            |
| pnn                                  | 8340.173            | 6239.980              | 2481.660             | 2481.660               | 946              |
| orthovisc_monovisc                   | 2235.463            | 1391.926              | 732.236              | 858.645                | 1288             |
| bdet                                 | NaN                 | NA                    | NaN                  | NaN                    | 0                |
| post_TLS_fusion                      | NaN                 | NA                    | 6283.426             | 6277.896               | 0                |
| tavr                                 | NaN                 | NA                    | 1724.636             | 1674.128               | 0                |



There does not appear to be any immediately obvious geographical trend. Some large metropolitan areas appear to show evidence of higher insurance payouts, but gaps in data make interpretation difficult. Let us try visualizing per capita income across MSAs:

```
per_capita_income_data <- read.csv("MSAs income per capita (csv).csv")
per_capita_income_data$Per.capita.personal.income.2018 = as.integer(per_capita_income_data$Per.capita.p
per_capita_income_data <- full_join(per_capita_income_data, msa_geo, by=c("i..Metropolitan.Statistical..
ggplot(data=per_capita_income_data) +
  geom_sf(data=us_geo,aes(geometry=geometry)) +
  geom_sf(aes(fill=Per.capita.personal.income.2018,geometry=geometry)) +
  coord_sf(xlim=c(-170,-70),ylim = c(20,70))
```

