Opportunity Insights Economic Tracker Data Documentation

last updated on 2021-04-23

1 Overview

This document provides an overview of the sources and processing applied to each data series within the Opportunity Insights Economic Tracker. The documentation is organized sequentially by series in the tracker, then broken down into categories of information describing each series, its source data, and our processing steps.

You can refer to additional documentation published by Opportunity Insights for complementary information:

- The Economic Tracker's **Data Dictionary** lists each data file and variable available for public use, with short descriptions of the contents of each variable.
- The **accompanying paper** provides detailed information about the methodology used to construct the series.

Please note that both the data and this data documentation is updated regularly and that the following information is subject to change.

2 Data Series

2.1 Consumer Spending

Summary: Aggregated and anonymized purchase data from consumer credit and debit card spending. Spending is reported based on the ZIP code where the cardholder lives, not the ZIP code where transactions occurred.

Data Source: Affinity Solutions

Update Frequency: Weekly

Date Range: January 13th until the most recent date available.

Data Frequency: Data is daily until the final two weeks of the series, and the daily data is presented as a 7 day lookback moving average. For the final two weeks of the series, the data is weekly and presented as weekly data points.

Index Period: January 4th - January 31st

Indexing Type: Seasonally adjusted change since January 2020. Data is indexed in 2019 and 2020 as the change relative to the January index period. We then seasonally adjust by dividing year-over-year, which represents the difference between the change since January observed in 2020 compared to the change since January observed since 2019. We account for differences in

the dates of federal holidays between 2019 and 2020 by shifting the 2019 reference data to align the holidays before performing the year-over-year division.

Geographies: National, State, County, Metro

Breakdowns:

- *By Industry.* Industries are constructed by grouping merchant codes that are used by Affinity Solutions to identify the category of merchant and merchant activity.
  - Apparel and General Merchandise
  - Entertainment and Recreation
  - Grocery
  - Health Care
  - Resturants and Hotels
  - Transportation
- *By Consumer Zip Code Income.* Transactions are linked to zip codes where the consumer lives and zip codes are classified into income categories based on measurements of median household income and population provided by the American Community Survey (2014 - 2018).
  - High Income (median household income greater than $78,000 per year)
  - Middle Income (median household income between $46,000 per year and $78,000 per year)
  - Low Income (median household income less than $46,000 per year)

Data masking: For the state-level breakdowns by income quartile and the county-level data, we mask locations with average daily spending $<$ $70,000 in January 2019. The raw data contains discontinuous breaks caused by entry or exit of credit card providers from the sample: counties with multiple structural breaks are dropped from the sample. Additionally, Affinity Solutions suppresses any cut of the data with fewer than five transactions. For more details refer to the accompanying paper.

Notes: We require at least 3 weeks of data in order to reliably identify and correct discontinuous breaks caused by entry or exit of credit card providers from the sample. The most recent 3 weeks of data are therefore marked 'provisional' and are subject to non-negligible changes as new data is posted. For breaks found prior to the last 3 weeks, we correct for it using a method outlined in the paper. Otherwise we substitute the national mean for more recent breaks while we gather enough data to implement the corrections outlined in the paper.

2.2 Small Business Revenue

Summary: Small business transactions and revenue data aggregated from several credit card processors. Transactions and revenue are reported based on the ZIP code where the business is located.

Data Source: Womply

Update Frequency: Weekly

Date Range: January 15th until the most recent date available.

Data Frequency:

- *National, State, Metro*: Daily, presented as a 7-day moving average
- *County*: Weekly, presented as a 6-day average Monday through Saturday, omitting Sunday

Indexing Period: January 4th - January 31st

Indexing Type: Seasonally adjusted change since January 2020. Data is indexed in 2019 and 2020 as the change relative to the January index period. We then seasonally adjust by dividing year-over-year, which represents the difference between the change since January observed in 2020 compared to the change since January observed since 2019. We account for differences in the dates of federal holidays between 2019 and 2020 by shifting the 2019 reference data to align the holidays before performing the year-over-year division. For series at the weekly frequency, we define weeks to run from Monday through Saturday and drop Sunday data. (See the Data Masking section below.) Weeks that span the end of the year are treated as the first week of the later year.

Geographies: National, State, County, Metro

Breakdowns:

- *Industry*, by NAICS supersector.
  - Education and Health Services
  - Leisure and Hospitality
  - Professional & Business Services
  - Retail and Transportation
- *Business Zip Code Income*. Transactions are linked to ZIP codes where the business is located and ZIP codes are classified into income categories based on measurements of median household income and population provided by the American Community Survey (2014 - 2018).
  - High Income (median household income greater than $78,000 per year)
  - Middle Income (median household income between $46,000 per year and $78,000 per year)
  - Low Income (median household income less than $46,000 per year)

Data Masking: The sample is restricted to firms with 30 or more transactions in a quarter and more than one transaction in 2 out of the 3 months. To reduce the influence of outliers, Womply excludes firms outside twice the interquartile range of annual firm revenue calculated within the sample. To preserve the privacy of firms, Womply imputes values for cells that contain fewer than 3 merchants.

For the county series, which is measured with a weekly frequency, we reduce the influence of imputation by dropping all data from Sundays (which are disproportionately likely to contain imputations). We drop any County x Week cell that contains imputed data within Monday-Saturday and we drop counties entirely if over 25% of their weeks contain imputed data.

We also exclude counties with a total average revenue of less than $250,000 or an average revenue of less than $10,000 during the indexing period (January 4-31, 2020). Additionally we omit spending categories for a small number of geographies that are extreme positive outliers, and we cap a small number of extreme negative outliers at 0 revenue.

Notes:

Small businesses are defined as those with annual revenue below the Small Business Administration's thresholds. Thresholds vary by 6 digit NAICS code ranging from a maximum number of employees between 100 to 1500 to be considered a small business depending on the industry.

County-level and metro-level data and breakdowns by High/Middle/Low income ZIP codes have been temporarily removed since the August 21st 2020 update due to revisions in the structure of the raw data we receive. We hope to add them back to the OI Economic Tracker soon.

2.3 Small Businesses Open

Summary: Number of small businesses open, as defined by having had at least one transaction in the previous 3 days.

Data Source: Womply

Update Frequency: Weekly

Date Range: January 15th until the most recent date available.

Data Frequency:

- *National, State, Metro*: Daily, presented as a 7-day moving average
- *County*: Weekly, presented as a 6-day average Monday through Saturday, omitting Sunday

Indexing Period: January 4th - January 31st

Indexing Type: Seasonally adjusted change since January 2020. Data is indexed in 2019 and 2020 as the change relative to the January index period. We then seasonally adjust by dividing year-over-year, which represents the difference between the change since January observed in 2020 compared to the change since January observed since 2019. We account for differences in the dates of federal holidays between 2019 and 2020 by shifting the 2019 reference data to align the holidays before performing the year-over-year division. For series at the weekly frequency, we define weeks to run from Monday through Saturday and drop Sunday data. (See the Data Masking section below.) Weeks that span the end of the year are treated as the first week of the later year.

Geographies: National, State, County, Metro

Breakdowns:

- *Industry*, by NAICS supersector.
  - Education and Health Services
  - Leisure and Hospitality
  - Professional & Business Services
  - Retail and Transportation
- *Business Zip Code Income*. Transactions are linked to ZIP codes where the business is located and ZIP codes are classified into income categories based on measurements of median household income and population provided by the American Community Survey (2014 - 2018).
  - High Income (median household income greater than $78,000 per year)
  - Middle Income (median household income between $46,000 per year and $78,000 per year)
  - Low Income (median household income less than $46,000 per year)

Data Masking: The sample is restricted to firms with 30 or more transactions in a quarter and more than one transaction in 2 out of the 3 months. To reduce the influence of outliers, Womply excludes firms outside twice the interquartile range of annual firm revenue calculated within the sample. To preserve the privacy of firms, Womply imputes values for cells that contain fewer than 3 merchants.

For the county series, which is measured with a weekly frequency, we reduce the influence of imputation by dropping all data from Sundays (which are disproportionately likely to contain imputations). We drop any County x Week cell that contains imputed data within Monday-Saturday and we drop counties entirely if over 25% of their weeks contain imputed data.

We also exclude counties with a total average revenue of less than $250,000 or an average revenue of less than $10,000 during the indexing period (January 4-31, 2020). Additionally we omit spending categories for a small number of geographies that are extreme positive outliers, and we cap a small number of extreme negative outliers at 0 revenue.

Notes: Small businesses are defined as those with annual revenue below the Small Business Administration's thresholds. Thresholds vary by 6 digit NAICS code ranging from a maximum number of employees between 100 to 1500 to be considered a small business depending on the industry.

County-level and metro-level data and breakdowns by High/Middle/Low income ZIP codes have been temporarily removed since the August 21st 2020 update due to revisions in the structure of the raw data we receive. We hope to add them back to the OI Economic Tracker soon.

2.4 Job Postings

Summary: Weekly count of new job postings, sourced from over 40,000 online job boards. New job postings are defined as those that have not had a duplicate posting for at least 60 days prior.

Data Source: Burning Glass Technologies

Update Frequency: Weekly

Date Range: January 17th until the most recent date available.

Data Frequency: Weekly data points, with each week ending on Friday.

Indexing Period: January 4th - January 31st

Indexing Type: Change relative to the January 2020 index period, not seasonally adjusted.

Geographies: National, State, Metro.

Breakdowns:

- *Industry*, by NAICS supersector.
  - Educational and Health Services
  - Financial Activities and Services
  - Leisure and Hospitality
  - Manufactoring
  - Professional and Business Services
- *Education Requirement*, by ONET Jobzone's Education Requirement Classification.
  - Minimal - Jobzone 1
  - Some - Jobzone 2
  - Moderate - Jobzone 3
  - Considerable - Jobzone 4
  - Extensive - Jobzone 5

Data Masking: In order to avoid extreme outliers, we calculate a cutoff of one standard deviation above the 97th percentile of the state-level data for each variable and mask values that exceed this threshold.

2.5 Employment

Summary: Number of active employees, aggregating information from multiple data providers. This series is based on firm-level payroll data from Paychex and Intuit, worker-level data on employment and earnings from Earnin, and firm-level timesheet data from Kronos.

Data Source: Paychex, Intuit, Earnin, Kronos

Update Frequency: Weekly

Date Range: January 15th 2020 until the most recent date available. The most recent date available for the full series depends on the combination of Paychex, Intuit and Earnin data. We extend the national trend of aggregate employment and employment by income quartile by using Kronos timecard data and Paychex data for workers paid on a weekly paycycle to forecast beyond the end of the Paychex, Intuit and Earnin data.

Data Frequency: Daily, presented as a 7-day moving average

Indexing Period: January 4th - January 31st

Indexing Type: Change relative to the January 2020 index period, not seasonally adjusted.

Geographies: National, State, County, Metro

To prevent the introduction of new Paychex clients from artificially creating noise in the employment series overtime, in the underlying raw data we suppress county x quartile x industry x firm size cells that both (i) experience a large anomalous change in employment and (ii) made up a large share of given wage quartile's total employment at any point in a county in the current year. For more details on the specifics of these thresholds see the appendix of the accompanying paper.

Breakdowns:

- *Wage.*
  - High Income (wage greater than $60,000 per year)
  - Middle Income (wage between $27,000 per year and $60,000 per year)
  - Low Income (wage less than $27,000 per year)
- *Industry*, by NAICS supersector.
  - Professional and Business Services
  - Education and Health Services
  - Retail and Transportation
  - Leisure and Hospitality
- *Industry*, by NAICS sector.
  - Retail

Data masking: As the employment series is a composite series, each of its component series have their own masking standards that in aggregate determine masking for the series.

In the Paychex series, we perform masking in order to avoid the introduction of new Paychex clients from artificially distorting a series through structural breaks in the underlying data. We define "influential cells" that are most sensitive to the introduction of new clients to the data and drop those "influential cells" that change significantly over the course of the year. We specifically denote county x wage quartile x industry x firm size bin cuts as an "influential cell" if either

- the county contains 100 or fewer unique county x quartile x industry x firm size cuts and that cut accounts for over 10% of employment in the corresponding county x wage quartile at any date in 2020 or,
- the county contains greater than 100 unique county x quartile x industry x firm size cuts and that cut accounts for over 5% of employment in the corresponding county x wage quartile at any date in 2020.

We then drop "influential cells" that record growth in employment exceeding 50% relative to January 2020 on any date, in order to omit trends likely arising due to changes in Paychex's client base rather than true employment changes.

In the Earnin series, we restrict the sample to workers who are active Earnin users with non-missing earnings and hours worked over the last 28 days and we exclude workers whose reported income over the prior 28 days is greater than \$50,000/13 (corresponding to an income of greater than \$50,000 annually).

In the Kronos and Intuit series, we do not make any sample restrictions.

Notes:

- For low income workers, the change in employment is calculated using Paychex and Earnin data. For medium and high income workers, the change in employment is calculated using Paychex and Intuit data.
- In order to provide closer to real time data, we forecast the most recent employment measures beyond those available in the combined Earnin, Intuit, and Paychex dataset alone. To do so, we leverage two sources of higher frequency data: Kronos timestamp data and the Paychex weekly pay cycle sample. Using this higher frequency data we forecast more recent changes in employment using a distributed lag model, constructed by regressing a given week's employment measure on the corresponding week's Kronos measure, as well as its current and 3 previous lagged weeks' Paychex weekly pay cycle measure. For more details, please refer to the appendix of the accompanying paper.

2.6 Unemployment Claims

Summary: Weekly unimployment insurance claims counts and rates (as a share of the 2019 labor force) for all states, as well as initial unemployment insurance claims for select counties where the data is publicly available.

Data Source: State-level and national statistics are reported by the U.S. Department of Labor.

The county-level series is only available for states whose respective state agencies publish county level data:

- Alabama: Alabama Department of Labor
- Arizona: Arizona Commerce Authority
- California: Employment Development Department of California
- Colorado: Colorado Department of Labor and Employment
- Georgia: Georgia Department of Labor
- Hawaii: Hawaii Department of Labor
- Idaho: Idaho Department of Labor
- Illinois: Illinois Department of Employment Security
- Indiana: Indiana Department of Workforce Development
- Iowa: State of Iowa
- Kentucky: Kentucky Center for Statistics
- Maryland: Maryland Department of Labor

- Massachusetts: Massachusetts Department of Unemployment Assistance
- Missouri: State of Missouri
- Nebraska: NEworks (Government of Nebraska)
- Nevada: Nevada Department of Employment; Training and Rehabilitation
- New York: New York State Department of Labor
- Ohio: Ohio Department of Job and Family Services
- Pennsylvania: Government of Pennsylvania
- Washington: Washington State Employment Security Department
- Wisconsin: Wisconsin Department of Workforce Development
- Wyoming: Wyoming Department of Workforce Services

Update Frequency: Weekly (where available, in the case of county-level data)

Date Range: January 18th until the most recent date available.

Data Frequency: Weekly data points, with each week ending on Saturday.

Note that county-level claims in California, Georgia, Kentucky, and Illinois are reported at the monthly level and imputed to weekly data points for the county-level series. For more information about the imputation methodology, see the **accompanying paper**

Indexing Period: No indexing applied, the published numbers directly report quantities.

Indexing Type: No indexing applied, the published numbers directly report quantities.

Geographies: National, State, County, Metro.

Breakdowns:

- *Initial Claims*
  - Regular Claims
  - PUA Claims
  - Combined Claims
- *Continued Claims*
  - Regular Claims
  - PUA Claims
  - PEUC Claims
  - Combined Claims

Data masking: No masking is performed by Opportunity Insights, but county-level data is subject to varying masking rules implemented by the state agencies that release the data. For more details, check with the relevant state agency for that state's particular masking rules.

Notes: Unemployment claims rates are calculated by dividing unemployment claims counts by the Bureau of Labor Statistics labor force estimates from 2019.

Under the CARES Act, all states provide 13 additional weeks of federally funded Pandemic Emergency Unemployment Assistance (PEUC) benefits to people who exhaust their regular state benefits. Under the Act, through the end of 2020, some people who exhaust all these benefits, and others who have lost their jobs for reasons arising from the pandemic but who are not normally eligible for UI in their state, are eligible for Pandemic Unemployment Assistance (PUA). "Combined Claims" are defined as the sum of regular, PUA and PEUC unemployment benefit claims.

2.7 Online Math Participation

Summary: Number of students using Zearn Math, a curriculum from the non-profit Zearn, among schools that already used Zearn Math in course instruction before the pandemic.

Data Source: Zearn

Update Frequency: Weekly, except during summer and winter school breaks.

Date Range: January 6th to May 3rd 2020. The data series is not being updated during the summer. Updates will resume during the fall semester.

Data Frequency: Weekly data points, with each week ending on Sunday.

Indexing Period: January 6th - February 7th

Indexing Type: Change relative to the January 2020 index period, not seasonally adjusted.

Geographies: National, States, County, Metro

To ensure privacy, the data we obtain are masked such that any county with fewer than two districts, fewer than three schools, or fewer than 50 students on average using Zearn Math is excluded. Where possible, masked county levels values are replaced by commuting zone means.

Breakdowns:

- *School Income.* Schools are classified by income based on the share of students in the school eligible for free and reduced lunch based on data provided by Zearn.
  - High Income (35.7% students are free and reduced lunch eligible)
  - Middle Income (56.9% students are free and reduced lunch eligible)
  - Low Income (80.4% students are free and reduced lunch eligible)

Data masking: Data is masked such that any county with fewer than two districts, fewer than three schools, or fewer than 50 students on average using Zearn Math during the period between January 6 and February 7 is excluded. Masked county level data is replaced with the commuting zone average so long as there are more than two school districts in the commuting zone or at least three schools in the commuting zone. If these condition are not met the county-level data remains masked. Additionally we exclude schools who did not have at least 5 students using Zearn Math for at least one week from January 6 to February 7.

2.8 Student Progress in Math

Summary: Number of lessons completed by students each week using Zearn Math, among schools that already used Zearn Math in course instruction before the pandemic.

Data Source: Zearn

Update Frequency: Weekly, except during summer and winter school breaks.

Date Range: January 6th to May 3rd 2020. The data series is not being updated during the summer. Updates will resume during the fall semester.

Data Frequency: Weekly data points, with each week ending on Sunday.

Indexing Period: January 6th - February 7th

Indexing Type: Change relative to the January 2020 index period, not seasonally adjusted.

Geographies: National, States, County, Metro

To ensure privacy, the data we obtain are masked such that any county with fewer than two districts, fewer than three schools, or fewer than 50 students on average using Zearn Math is excluded. Where possible, masked county levels values are replaced by commuting zone means.

Breakdowns:

- *School Income.* Schools are classified by income based on the share of students in the school eligible for free and reduced lunch based on data provided by Zearn.
  - High Income (35.7% students are free and reduced lunch eligible)
  - Middle Income (56.9% students are free and reduced lunch eligible)
  - Low Income (80.4% students are free and reduced lunch eligible)

Data masking: Data is masked such that any county with fewer than two districts, fewer than three schools, or fewer than 50 students on average using Zearn Math during the period between January 6 and February 7 is excluded. Masked county level data is replaced with the commuting zone average so long as there are more than two school districts in the commuting zone or at least three schools in the commuting zone. If these condition are not met the county-level data remains masked. Additionally we exclude schools who did not have at least 5 students using Zearn Math for at least one week from January 6 to February 7.

2.9 COVID-19 Infections

Summary: The daily count and rate per 100,000 people of confirmed COVID-19 cases, deaths or tests performed.

Data Source: The Centers for Disease Control and Prevention

Update Frequency: Daily

Date Range: January 22th until the most recent date available.

Data Frequency: Daily, presented as a 7-day moving average

Indexing Period: No indexing applied, the published numbers directly report quantities.

Indexing Type: No indexing applied, the published numbers directly report quantities.

Geographies: National, State, Country, Metro

Note that testing counts and rates are only available at the national and state level, not at the county or metro levels.

Breakdowns:

- *New* Cases, Deaths, or Tests (presented as a 7-day moving average)
- *Total* Cases, Deaths, or Tests

Data masking: No masking is performed by Opportunity Insights.

2.10 COVID-19 Vaccinations

Summary: Percentage of the population who have received one or more doses of any COVID-19 vaccine.

Data Source: The Centers for Disease Control and Prevention

Update Frequency: Daily

Date Range: February 24th 2021 until the most recent date available.

Data Frequency: Daily, presented as a 7-day moving average for new vaccinations

Indexing Period: No indexing applied, the published numbers directly report quantities.

Indexing Type: No indexing applied, the published numbers directly report quantities.

Geographies: National, State

Breakdowns:

- *New Vaccinations* Percent of population newly vaccinated with at least one vaccine dose
- *Total Vaccinations* Percent of population in total vaccinated with at least one vaccine dose

Data masking: No masking is performed by Opportunity Insights.

Notes: CDC data published prior to the 24th of February 2021 used a different methodology to assign vaccinations to the state where they were administered, producing numbers that are not directly comparable to those published after February 24th.

2.11 Time Outside Home

Summary: Time spent away from home, estimated using cellphone location data from Google users who have enabled the Location History setting.

Data Source: Google COVID-19 Community Mobility Reports, American Time Use Survey

Update Frequency: When released by Google, typically every 4-7 days.

Date Range: February 24th until the most recent date available.

Data Frequency: Daily

Indexing Period: January 3rd to February 5th

Indexing Type: Change relative to the January 2020 index period, not seasonally adjusted.

Geographies: National, State, County, Metro

Breakdowns:

- Time Away From Home
- Retail and Restaurants
- Transit
- Parks
- Grocery
- Workplace

Data masking: Google does not release data for geographies where their internal quality and privacy thresholds are not met. Therefore some geographic areas are omitted from the series for certain breakdowns and certain dates.

Notes: When data is missing for 1 or 2 consecutive days we linearly interpolate the missing values and construct the 7 day moving average including these interpolated values. If data is missing for 3 or more consecutive days, the corresponding 7 day moving average is also recorded as missing whenever it overlaps with the missing data.

Time Away From Home is calculated by multiplying the mean time spent inside home from the American Time Use Survey by the percent change in time spent at residential locations reported by Google. For more information about this imputation, see the **accompanying paper**.

2.12 Policy Milestones

Summary: Key state-level policy dates relevant for changes in other series trends and values. Includes start and end of stay at home order dates, public school closure dates, and non-essential business closure and re-opening dates.

Data Source(s): New York Times, MCH Strategic Data, the Institute for Health Metrics and Evaluation, and local news and government sources.

Update Frequency: Monthly

Geographies: State